Combinando datos GPS en contextos offline y online para predecir el tiempo de llegada de los buses en sistemas BRT

Andrea Paola Martínez Marín

Universidad del Norte

Departamento de Ingeniería Civil, División de Ingenierías

Barranquilla, Colombia

2020

Combinando datos GPS en contextos offline y online para predecir el tiempo de llegada de los buses en sistemas BRT

Andrea Paola Martínez Marín

Tesis de investigación presentada como requisito parcial para optar al título de:
Magister en Ingeniería Civil

Directores:
Julián Arellana Ph.D.
Elías Niño Ph.D.

Grupo de Investigación:
TRANVIA

Universidad del Norte

Departamento de Ingeniería Civil, División de Ingenierías

Barranquilla, Colombia

2020

RESUMEN

Obtener predicciones precisas de la hora de llegada de los buses en tiempo real es un elemento vital tanto para el control de las operaciones de los autobuses como para los sistemas de información de pasajeros. Los avances tecnológicos asociados a los sistemas de transporte público han permitido la recopilación y difusión de grandes cantidades de información sobre las características del tráfico en tiempo real, que abren la puerta para el desarrollo e implementación de diferentes metodologías de predicción que permitan realizar estas predicciones. El objetivo de esta investigación es formular diversos modelos basados en datos históricos y en tiempo real para predecir la hora de llegada de los buses a las estaciones de un sistema troncal de transporte masivo. Usando información proveniente de los servicios troncales del sistema BRT de la ciudad de Pereira, Colombia, se evaluaron los métodos de velocidad promedio, regresión lineal, redes neuronales artificiales (ANN), máquinas de vectores de soporte (SVM), regresión Ridge, regresión Lasso y técnicas de inferencia bayesiana (asimilación de datos). La comparación de los resultados obtenidos, luego de realizar la validación de cada uno de los métodos considerados, indicó que el modelo que considera técnicas de inferencia bayesiana tiene un mejor ajuste con respecto a los datos observados, y que además se puede implementar fácilmente en aplicaciones de predicción online, por su carácter dinámico y eficiencia computacional. Los resultados también sugieren que, para el contexto estudiado, la inclusión del tiempo para llegar a la parada como única variable en el modelo propuesto es suficiente para obtener predicciones precisas, evitando la necesidad de medir o inferir otras variables explicativas que influyen en la hora de llegada a los paraderos, las cuales son comúnmente consideradas.

Palabras clave: transporte público, ITS, tiempo de viaje, horas de llegada a paraderos, inferencia bayesiana, asimilación de datos.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1 INTRODUCCIÓN

La incorporación de las tecnologías de información y comunicación en los sistemas de gestión e infraestructura del transporte ha derivado en el desarrollo de los Sistemas Inteligentes de Transporte (ITS por sus siglas en inglés, Intelligent Transportation System). Los ITS nacieron como una solución tecnológica para mejorar la gestión del tránsito y resolver el problema de congestión vehicular que se ha generado a partir del incremento de la demanda por transporte, causado por mayores niveles de desarrollo de las ciudades y el crecimiento demográfico. Los avances tecnológicos asociados a estos sistemas han facilitado la recopilación y difusión de información sobre el tráfico en tiempo real, lo que ha permitido que los planificadores y operadores de los sistemas de transporte analicen y estudien la manera como dichos sistemas se comportan. Además, ha abierto la puerta para que numerosos investigadores propongan soluciones para los problemas de tráfico a partir de la predicción de variables tales como flujos, tiempos de viaje entre orígenes y destinos, y la hora de llegada a los paraderos.

Hablando específicamente del transporte público, el desarrollo de estrategias eficientes y efectivas que buscan mejorar la operación del sistema de transporte son la base en la cual se fundamenta los Sistemas Avanzados de Transporte Público (APTS). Estos persiguen la búsqueda de la integración de todos los agentes viales propiciando el vínculo entre la infraestructura de transporte, el vehículo, el conductor, y el pasajero. Para lograrlo se valen de herramientas como el procesamiento de información, navegación y localización, reconocimiento de imágenes, redes y otras técnicas.

Para garantizar un transporte sostenible y de calidad en las zonas urbanas, es importante proporcionar un servicio de transporte público que provea altos niveles de servicio. El nivel de servicio en estos casos depende de varios atributos como lo son: el tiempo de espera en la parada, el tiempo de viaje en el autobús, la ocupación del vehículo, la limpieza del vehículo, la hospitalidad del conductor y la comodidad de los autobuses (Dell'Olio et al., 2011). Un componente especial para ofrecer un buen nivel de servicio es la información de calidad, fácilmente comprensible y confiable. Una de las variables más relevantes y que más se ha estudiado en la literatura es el tiempo de viaje en tiempo real (Zhou et al., 2014). Este tipo de información puede estar disponible para los pasajeros de los sistemas de transporte público

a través de diversos medios de información, como pantallas de información instaladas en las paradas de buses y en dispositivos móviles, como teléfonos inteligentes y tabletas. Específicamente se pueden encontrar sistemas de información que comparten información sobre los tiempos de llegada y salida de los buses a las paradas.

Se han identificado muchos efectos potenciales asociados a los sistemas de información en tiempo real en el transporte público (Dziekan, 2008). Entre ellos se destacan una serie de efectos positivos: disminución del tiempo de espera percibido en las paradas, percepción positiva asociada a una menor incertidumbre, mayor facilidad de uso del sistema y mayor sensación de seguridad, mayores disposiciones a pagar, mejor aprovechamiento del tiempo de espera y viajes más eficientes, impacto en la elección del modo de transporte, y una mejora general en la imagen del transporte público. Una condición es que la información sea relevante y confiable, ya que de lo contrario podría tener un efecto negativo.

La posibilidad de obtener predicciones precisas de la hora de llegada de los buses a las paradas en tiempo real es un elemento vital tanto para el control de las operaciones de los autobuses como para los sistemas de información de pasajeros (Zheng et al., 2012). En el caso de las agencias de tránsito, con esta información disponible estas pueden administrar y operar sus sistemas de manera más efectiva con despacho y programación en tiempo real, monitoreando la ejecución del cronograma y evaluando la eficiencia operativa. De igual manera, ante alguna interrupción del servicio, la información en tiempo real permite tomar las acciones de control o corrección necesarias, como lo son aumentar o disminuir la velocidad de operación o indicar tiempos de permanencia más largos en algunas paradas (Chen et al., 2004; Yu et al., 2011). Incluso abre la posibilidad a los operadores de construir herramientas de planificación de viaje. En cuanto a los pasajeros, estos sistemas deben ayudar a los usuarios a tomar decisiones antes del viaje y durante la ruta, lo que les permite planificar su tiempo de viaje y ahorrar tiempo de espera en las paradas de bus (Wang et al., 2014). Estos efectos mejoran la percepción y popularidad de los autobuses, lo que ayuda a promover el uso de los sistemas de transporte público en pro de la sostenibilidad. Sin embargo, todavía es una tarea difícil predecir la hora exacta de llegada del autobús debido a las condiciones dinámicas del tráfico.

Aunque se han desarrollado numerosos productos y sistemas para predecir datos en tiempo real en transporte público, todavía existen desviaciones entre las predicciones y el tiempo de viaje real. Estas desviaciones suelen deberse a varios factores estocásticos. Las operaciones de los vehículos de transporte público se ven perturbadas con frecuencia por la congestión en la ruta de servicio que además es variable entre diferentes momentos del día, la interacción con otros tipos de vehículos, variaciones en las demandas, tiempos de parada variables en las estaciones, e incluso el clima (M. Chen et al., 2004). Las discrepancias entre la predicción y la hora real de llegada del autobús a la parada son particularmente comunes en los sistemas de transporte público de autobuses que utilizan la misma infraestructura vial que otros modos de transporte. También cabe destacar que, los tiempos de viaje del autobús en diferentes períodos de tiempo, por ejemplo, durante las horas pico, no pico, o durante los fines de semana, pueden variar significativamente.

El tiempo estimado de llegada del bus a la estación de destino se calcula en función del conocimiento de la ubicación actual del vehículo y del tiempo de viaje estimado que necesita el bus desde la ubicación actual hasta la estación de destino. Hoy en día se obtienen enormes cantidades de datos sobre la ubicación de los vehículos a través del sistema de posicionamiento global (GPS) que permite determinar la ubicación actual y obtener muestras de velocidades de viaje pasadas (Arellana et al., 2014; Čelan & Lep, 2017) .

Aprovechando la disponibilidad de este tipo de información, el objetivo de esta investigación es formular una variedad de modelos basados en datos históricos y datos en tiempo real cuyo propósito será la predicción de los tiempos de llegada de los buses a las estaciones. Se emplearon los métodos de velocidad promedio, regresión lineal, redes neuronales artificiales (ANN), máquinas de vectores de soporte (SVM), regresión Ridge, regresión Lasso y además un modelo formulado con técnicas bayesianas propuesto para el desarrollo de esta tesis. Finalmente se realizó una comparación de los resultados obtenidos para establecer recomendaciones para la predicción de las horas de llegada de los buses y cuál enfoque obtiene los mejores resultados.

1.1. Objetivos

1.1.1. Objetivo general

Formular y estimar diversos modelos basados en datos históricos y en tiempo real para predecir la hora de llegada de los buses a las estaciones de un sistema troncal de transporte masivo.

1.1.2. Objetivos específicos

- Identificar las variables que caracterizan el comportamiento del tiempo de llegada de los buses a los paraderos de un sistema troncal a partir de información GPS.
- Usar información histórica asociada a la operación de un sistema troncal para estimar diferentes familias de modelos que permitan predecir la hora de llegada de los buses.
- Formular modelos de inferencia bayesiana para predecir la hora de llegada de los buses en un contexto online.
- Comparar los resultados de predicción de los modelos offline y online usando diferentes medidas de desempeño.

1.2. Alcance y limitaciones

La metodología desarrollada está pensada para predecir la hora de llegada de los buses a las estaciones de un sistema troncal de transporte masivo mediante la información GPS de los vehículos en operación. En esta oportunidad no fueron incluidas las rutas periféricas del sistema de transporte, debido a que éstas no cuentan con paraderos de ascenso y descenso de pasajeros definidos.

De igual forma, cabe aclarar que los modelos propuestos y desarrollados a continuación analizan únicamente la información GPS de los buses de una de las dos empresas que operaban el sistema BRT, cuando fue tomada la información. Esto debido a que los vehículos de la flota de la otra empresa contaban con dispositivos GPS incorporados.

Los modelos planteados se encuentran basados en la información histórica de los registros GPS de los buses que se tiene disponible. En caso tal que las condiciones operacionales del sistema presentan un cambio significativo, se debe recolectar nuevos datos y volver a reentrenar el modelo estimado con las nuevas circunstancias.

Además, es importante mencionar la segmentación de la información propuesta en esta investigación para realizar los diferentes modelos de predicción se adaptan al presente caso de estudio, ya que cada sistema de transporte cuenta con sus propias características operacionales tales como flujos de vehículos, si cuenta con un carril exclusivo para el tránsito de los buses, demanda de pasajeros; y otros factores como lo son los diferentes meses del año y las temporadas de vacaciones, que afectan directamente los criterios para segmentar la información para realizar una predicción adecuada bajo estas condiciones.

Por otro lado, no fue tenido en cuenta el efecto que puede tener en el error de la estimación dada la cantidad de información histórica empleada para el modelo, es decir, la variabilidad que puede presentarse en el acierto de los modelos de predicción debido al número de días de registros GPS de los buses utilizados para estimarlos.

2 ANTECEDENTES

Los sistemas de transporte de autobuses de tránsito rápido (BRT por sus siglas en inglés) son aquellos donde los buses se separan físicamente de los automóviles a través de la provisión de carriles exclusivos, con vías de autobús y estaciones típicamente alineadas con el centro de la carretera, donde el cobro de las tarifas se hace antes de abordar (Basso et al., 2019).

Asimismo, al estar segregado de los demás vehículos que circulan en las vías, el impacto de otros éstos en el tiempo de viaje de los buses dentro del sistema BRT es bajo. Además, los vehículos generalmente están equipados con un sistema de posicionamiento global (GPS), que puede proporcionar soporte técnico y datos para predecir el tiempo de viaje y de llegada a las estaciones de los buses en el sistema BRT (Chen et al., 2012).

Numerosas investigaciones, desde finales de la década de los 90, han estudiado diversidad de métodos para la predicción de los tiempos de llegada de los buses a las paradas o estaciones de los sistemas de transporte público de pasajeros. El propósito de dichas investigaciones ha sido extraer información sobre el funcionamiento de los buses de los sistemas de monitoreo vehicular (Lin y Zeng, 1999). De igual manera se han estudiado los efectos de los factores que influyen indirectamente en la velocidad del autobús y por lo tanto en el tiempo de viaje de los mismos, como lo son el tiempo de parada (en estación usualmente para el ascenso y descenso de pasajeros), el clima, las condiciones del tráfico, la demanda de viajes, el distanciamiento entre buses, tipos de buses, métodos de pago, entre otros (Čelan & Lep, 2020).

El tiempo estimado de llegada de los buses a la parada de destino se calcula en función de sus ubicaciones actuales y el tiempo de viaje estimado para que cada uno de los buses llegue de dicha ubicación a la parada de destino. La información sobre la ubicación de los vehículos puede ser obtenida de forma masiva a través de los sistemas de posicionamiento global (GPS) (Elragal & Raslan, 2014), que además permiten recopilar muestras históricas de la velocidad y el tiempo (Arellana et al., 2014).

En un intento por distinguir los distintos tipos de aportes y metodologías planteadas a lo largo de los últimos años para la estimación o predicción de los tiempos de viaje, algunos autores

han propuesto algunas clasificaciones de los modelos. Altinkaya y Zontul (2013) clasificaron los modelos en cuatro categorías: modelos basados en datos históricos, modelos estadísticos, modelos de filtrado de Kalman y modelos de aprendizaje de máquinas (*Machine learning*).

Por su parte, Mori et al. (2015) hicieron un artículo de revisión en el cual consideraron clasificaciones basadas en varios aspectos: la fuente y formato de la información (por ejemplo, tipos de sensores), contexto de estudio (autopistas o arterias urbanas) y el propósito (estimación o predicción de los tiempos de viaje). El objetivo de los modelos de predicción es pronosticar el tiempo de viaje para una trayectoria que comenzará en el momento en que se haga la predicción (presente) o en el futuro. Para hacerlo se incluye la variable del tiempo e información sobre el tráfico o el contexto en el presente, considerando además datos del pasado.

Dentro de los modelos de predicción, objetivo de interés para el desarrollo de esta tesis, hicieron a su vez una clasificación en función del tipo de modelo aplicado. En este caso se distinguen los modelos ingenuos o "naive", los modelos basados en datos y los modelos basados en teoría del tráfico.

Los enfoques ingenuos son los métodos más simples de predicción del tiempo de viaje. La principal ventaja de estos métodos es que son rápidos, computacionalmente hablando, y fáciles de implementar. Sin embargo, suelen estar basados en supuestos que no se cumplen en muchas situaciones, como por ejemplo que las condiciones sobre un tramo de vía en particular permanecerán constantes en el tiempo (Van lint, 2004). Entre estos métodos simples se encuentran los pronósticos por velocidad promedio, el cual se vale de las velocidades comerciales medias monitoreadas por los GPS. Los promedios se pueden calcular para cada ruta de autobús o sobre segmentos de longitud arbitraria, además de dividirse en intervalos de tiempo de diferente duración (Cortés et al., 2011).

En cuanto a los enfoques basados en datos, estos desarrollan relaciones entre variables dependientes e independientes. Dentro de ellos se distinguen los modelos paramétricos en los cuales se predefine el conjunto de parámetros que deben estimarse para establecer una relación entre las variables. Dichos parámetros se calculan utilizando datos. Los modelos paramétricos más populares son el análisis de series de tiempo y las técnicas de regresión. Es

importante mencionar que en el caso de los modelos de regresión, todas las variables exploratorias deben ser independientes entre ellas por lo que su usó en el ámbito del transporte puede verse limitado dado que las variables suelen estar interrelacionadas (Bae, 1997). Sin embargo, el efecto de multicolinealidad puede ser reducido usando enfoques de regresiones robustas como la regresión Ridge (Haworth et al., 2014) o la regresión Lasso (Kamarianakis, Shen, y Wynter, 2012).

Igualmente, dentro de los enfoques basados en datos también se destacan los modelos no paramétricos en los que la estructura del modelo no está predefinida y, por lo tanto, la relación entre las variables dependientes e independientes se obtiene de los propios datos. Los modelos no paramétricos comúnmente utilizados en la literatura de predicción del tiempo de viaje son enfoques de aprendizaje de máquinas entre los que se destacan las redes neuronales artificiales (ANN por sus siglas en inglés) y las máquinas de vectores de soporte (SVM por sus siglas en ingles). La ventaja que ofrecen este tipo de métodos es que pueden capturar relaciones no lineales complejas, las cuales pueden darse comúnmente en vías congestionadas, o en situaciones causadas por incidentes de tránsito y contextos urbanos (Mori et al., 2015).

Por último, los enfoques basados en la teoría del tráfico generalmente se enfocan en recrear las condiciones del tráfico en los intervalos de tiempo futuros y luego derivar los tiempos de viaje (Van lint, 2004). Estos pueden ser muy útiles para estudiar contextos con congestión o zonas urbanas, sin embargo, suelen ser modelos complicados y computacionalmente costosos, lo que dificulta su aplicación en redes enteras, especialmente para fines de estimación o predicción en tiempo real.

A continuación, se muestran algunas de las investigaciones más destacadas en cada uno de los enfoques descritos.

D'Angelo et al. (1999) utilizó un modelo de series de tiempo no lineal para predecir el tiempo de viaje de un corredor en una carretera. Comparó dos casos: en el primer modelo sólo consideró la velocidad, mientras que el segundo modelo usó datos de velocidad, ocupación y flujo para predecir el tiempo de viaje. Finalmente determinaron que el modelo que solo consideraba la velocidad era mejor que el modelo de predicción multivariable.

Sun et al. (2007) usaron un algoritmo de predicción de la hora de llegada del autobús que combina los datos del Sistema de posicionamiento global (GPS) con velocidades de viaje promedio de segmentos de ruta individuales, teniendo en cuenta la velocidad de viaje histórica y las variaciones temporales y espaciales de las condiciones del tráfico.

Patnaik et al. (2004) propusieron modelos de regresión multilineal para estimar los tiempos de llegada de los autobuses utilizando los datos recopilados por el contador automático de pasajeros (APC). Consideraron la distancia, el número de paradas, los tiempos de parada, el número de pasajeros que suben y bajan y las condiciones climáticas como variables independientes para realizar la predicción.

Jeong & Rilett (2004) propusieron un modelo de redes neuronales artificiales (ANN) para predecir los tiempos de llegada de los autobuses y demostraron que su rendimiento es superior que el de modelos de regresión multilineal basados en datos históricos. En sus modelos consideraron como entrada para la predicción de la congestión, el cumplimiento de los horarios de la programación de los buses y los tiempos de permanencia en las paradas.

Chien y Kuchipudi (2003) usaron filtros de Kalman para la predicción del tiempo de viaje usando datos históricos y con datos en tiempo real recopilados en la autopista NYST del estado de Nueva York. El tiempo de viaje promedio de los vehículos estudiados en cada intervalo de tiempo se trata como el valor real para predecir el tiempo de viaje en el próximo período de tiempo. Consideraron información sobre los tiempos de viaje de los vehículos, velocidades promedio, desviaciones estándar de tiempos de viaje y volúmenes de tráfico entre lectores sucesivos (RST- *Road side terminals*)

Cathey y Dailey (2003) plantearon una prescripción general para la predicción de la llegada/salida de los buses. En dicha prescripción identificaron los tres componentes necesarios para realizar la tarea de predicción: un rastreador, un filtro y un predictor. Estas actividades son necesarias para tomar los datos de ubicación automática del vehículo (AVL), posicionar cada vehículo en el espacio y el tiempo y luego predecir la llegada/salida en una ubicación seleccionada. En cuanto al componente del filtro propusieron un filtro de Kalman para obtener estimaciones óptimas de ubicación y velocidad. Para representar el estado

dinámico instantáneo de un vehículo, seleccionaron un espacio de estado tridimensional que incluye la distancia recorrida, la velocidad del bus y la aceleración.

En la propuesta de Chen et al. (2004), un modelo de redes neuronales (ANN) proporciona un conjunto de estimaciones iniciales del tiempo de viaje desde el origen hasta cada punto aguas abajo utilizando información de los sistemas de conteo automático de pasajeros (APC). Se basan además en una serie de variables de entrada (día de la semana, hora del día, clima y segmento) relacionadas con cada viaje. Cuando el autobús llega al segundo punto, se registra el tiempo de viaje real entre el origen y este punto de tiempo, y un filtro de Kalman actualiza los tiempos de viaje estimados desde la parada actual hasta todos los puntos de tiempo aguas abajo en función de esta información.

Bin et al. (2006) utilizaron el tiempo de viaje en cada tramo de los buses anteriores y del bus actual para estimar las condiciones del tráfico de la red y desarrolló los modelos de predicción, basados en el método de máquina de vectores de soporte (SVM, por sus siglas en inglés) con tres variables de entrada: información del segmento, tiempo de viaje del segmento actual, y el último tiempo de viaje del siguiente segmento. Estimaron modelos diferentes para distintas horas del día y condiciones climáticas.

Por su parte, Zheng et al. (2012) presentaron un modelo de predicción iterativo formulado SVM para determinar el tiempo de llegada en la parada siguiente aguas abajo y un modelo de predicción de velocidad promedio del autobús para los segmentos posteriores. Luego, los dos modelos de predicción se utilizan para pronosticar de manera recursiva los tiempos de llegada en múltiples paradas aguas abajo. Además, para mejorar la precisión de la predicción, se desarrolla un algoritmo dinámico basado en filtros de Kalman.

También, Deng, He, y Zhong (2013) propusieron un modelo en el cual se estableció una red bayesiana entre el estado del tráfico vehicular y el tiempo de viaje en autobús para obtener los parámetros de la red mediante el entrenamiento de datos históricos. Luego utilizaron la distribución conjunta de ambas variables para predecir el tiempo de viaje en autobús.

Wang et al. (2014) propusieron un enfoque que combina datos históricos e información en tiempo real en un modelo de dos fases. En primer lugar, el modelo de redes neuronales de función de base radial (RBFNN) se utiliza para aprender y aproximar la relación no lineal en

datos históricos. Las variables históricas consideradas fueron el tiempo de viaje, el tiempo de parada, la distancia, la cantidad de pasajeros que suben y bajan, el retraso del bus y la congestión. Luego, en la segunda fase, introdujeron un método en línea para ajustar a la situación real utilizando información real para modificar el resultado predicho por las RBFNN en la primera fase. En contraste, estimaron modelos de regresión lineal múltiple, redes neuronales BP y RBFNN sin ajuste en línea. Los resultados muestran que el enfoque con RBFNN y el ajuste en línea tiene un mejor rendimiento de predicción.

Por su parte, Xin y Chen (2016) se concentraron en la predicción del tiempo de los buses en las paradas. Su propuesta estuvo basada en el método de los K vecinos más cercanos en (kNN por sus siglas en inglés). Para la estimación de los tiempos de parada en las estaciones aguas abajo utilizando datos históricos de GPS del bus en periodos similares. El modelo propuesto se puede utilizar en la práctica sin necesidad de ajustes según el estilo del autobús, la forma de la parada y también sin necesidad de predecir el número de pasajeros que subirán y bajarán.

Kumar et al. (2017) desarrollaron un método de predicción que considera variaciones temporales y espaciales en el tiempo de viaje basado en los filtros de Kalman. Realizaron un análisis kNN para encontrar un viaje histórico que sea similar al viaje actual. La idea principal de tal análisis es separar los datos, basándose en algunas similitudes entre los patrones de viaje. A partir de los resultados, se encontró que el método propuesto fue capaz de funcionar mejor que los métodos de promedio histórico, regresión, redes neuronales y los métodos que consideraron variaciones temporales o espaciales por sí solas. Este estudio es uno de los primeros intentos de caracterizar y predecir la evolución del tiempo de viaje en autobús a lo largo del tiempo y el espacio mediante la reformulación de la ecuación básica de conservación de los vehículos, con la velocidad como variable de estado.

Yu et al. (2017) propusieron realizar la estimación de los tiempos de viaje de los buses usando modelos de supervivencia de tiempo de falla acelerado. Este estudio empleó datos de headway de una ruta de autobús que transita en el campus de la Universidad Estatal de Pennsylvania. El modelo de supervivencia propuesto fue comparado con un modelo tradicional de regresión lineal a través de medidas de error. Los resultados obtenido en el estudio sugieren que los modelos de supervivencia se pueden utilizar para brindarle a los

usuarios del sistema información de tiempo de viaje en autobús de mayor calidad, mejorando la confiabilidad del servicio prestado.

Hou & Edara (2018) proponen dos modelos de aprendizaje para predecir el tiempo de viaje en carretera, los cuales son la memoria larga a corto plazo (LSTM por sus siglas en inglés) y la red neuronal convolucional (CNN por sus siglas en inglés). El caso de estudio fue la red de transporte de la ciudad de Saint Louis, Missouri. También se desarrollaron otros dos modelos de máquina de aprendizaje para realizar la comparación: bosques aleatorios (RF por sus siglas en inglés) y máquinas de aumento de gradiente (GBM por sus siglas en inglés). Los resultados del estudio muestran que el aprendizaje profundo puede proporcionar una predicción precisa tanto para las condiciones de tráfico congestionado como no congestionado, capturando con éxito la dinámica del tráfico de incidentes inesperados o eventos especiales.

Xu et al. (2019) presentaron un método de estimación del tiempo de viaje basado en aprendizaje de máquinas usando los datos de vehículos equipados con GPS. Utilizaron un algoritmo de redes neuronales artificiales (ANN) luego de normalizar la información y agrupar los datos considerando las variaciones espaciales y temporales. Además, calcularon una suma ponderada de los resultados de la estimación del tiempo de viaje de varias trayectorias para representar mejor el tiempo de viaje del segmento en un intervalo.

Ma et al. (2019) realizaron una comparación entre los métodos de ANN, SVM y kNN con el propósito de elegir cual podría usarse como modelo básico en la estimación de los tiempos de viaje. Luego, determinaron que los patrones de viaje en autobús son bastante diferentes en condiciones de tráfico normales y anormales por lo que separaron los registros de autobuses irregulares basados en registros históricos. Para la predicción de tiempos de viaje utilizaron una combinación de bases de datos de taxis y autobuses en tiempo real, separando además la estimación del tiempo de tránsito de la estimación del tiempo de parada en estación. Para la estimación de los tiempos de parada consideraron, además de la información histórica del GPS, la demanda de pasajeros.

Tang et al. (2019) presentaron un modelo de arquitectura profunda, que utiliza codificadores automáticos de eliminación de ruido dispersos como bloques de construcción, para aprender

las representaciones de características para la estimación del tiempo de viaje. Se tiene en cuenta tanto las características geográficas como las características contextuales, y considera la correlación espacial de los segmentos de carreteras adyacentes. El caso de estudio fue aplicado a la red de carreteras de la ciudad de Bejing, China, usando información sobre las trayectorias GPS de taxis. Los resultados de los modelos de arquitectura profunda se compararon con un modelo de red neuronal de propagación trasera, obteniendo mejores resultados.

He et al. (2020) propusieron un modelo que predecía el tiempo de viaje y los tiempos de espera en los puntos de transbordo de los buses mediante el método Marco de Coalización de Segmento Centrado en Patrón de Tráfico (TP-SCF por sus siglas en inglés), el cual está pasado en patrones dispares aprendidos de las condiciones del tráfico en diferentes segmentos de rutas de buses. Este consta de una etapa de entrenamiento y otra de predicción. Los componentes del tiempo de viaje se predicen utilizando los modelos LSTM para cada uno de los de los grupos establecidos con condiciones de tráfico similares, mientras que los componentes del tiempo de espera se estiman en función de los registros históricos de tiempo de llegada del bus.

Dhivya Bharathi et al. (2020) presentaron dos metodologías para realizar la predicción del tiempo de llegada que usan los conceptos de los análisis de las series de tiempo: un modelo clásico estacional auto regresivo con la inclusión de efectos no estacionarios; y un enfoque lineal no estacionario auto regresivo. Los autores incorporaron en los modelos de predicción un análisis detallado de las distribuciones marginales de los datos; asimismo, propusieron un algoritmo de predicción del tiempo de viaje por adelantado de varias secciones para facilitar la implementación en tiempo real.

Recientemente, Comi et al. (2020) estudiaron los tiempos de viaje de los buses y su naturaleza sistemática y fluctuante por medio de métodos de series de tiempo analizando la información de monitoreo de los buses (AVL) en la ciudad de Lviv en Ucrania. Encontraron que la tendencia y la estacionalidad explican gran parte de la variabilidad en los tiempos de viaje.

A pesar de todos los métodos propuestos en la literatura, aún muchos de los ATIS más recientes se basan en los modelos más simples para realizar la predicción y estimación del

tiempo de viaje. Esto sucede porque los modelos más complejos generalmente no se adaptan adecuadamente a los requisitos prácticos de implementación, como lo son tratar con datos ruidosos o faltantes, hacer predicciones para una red completa, usar datos de diferentes fuentes, entre otros (Mori et al., 2015).

Luego de realizar la revisión de la literatura se llegó a la conclusión de que con el advenimiento de la era del big data, es posible plantear modelos eficientes que ignoren la relación entre el tiempo de viaje y los factores de influencia que afectan su variabilidad. Lo que realmente se necesita poder estudiar es cómo dicho tiempo de viaje varía y cuál es su "regla de cambio". Como es el caso, por ejemplo, de la estimación de tiempos de parada usando series de tiempo basadas en datos históricos. Una ventaja que ofrecen este tipo de modelos es que se pueden utilizar en la práctica sin necesidad de ajustar todas las variables influyentes que serán diferentes dependiendo de cada caso. Aunque el entrenamiento de los modelos basados en datos (especialmente, los no paramétricos) pueden ser computacionalmente complejos, una vez que se construye el modelo son muy útiles para realizar predicciones y estimación de forma online.

También cabe resaltar que, aunque se han publicado muchas investigaciones relacionadas con la predicción de los tiempos de viaje, hacen falta estudios comparativos completos que evalúen la bondad de diferentes modelos. En este momento, no hay suficiente información para elegir un "mejor método" para la estimación o predicción del tiempo de viaje. Además, en los pocos casos en los que se hacen comparaciones, suelen utilizarse los modelos más simples e ingenuos y no suelen comparar sus métodos con otros tipos de modelos avanzados. Aunque existe una vasta literatura en torno a este tema, los métodos han sido evaluados en diferentes contextos de tráfico y ubicaciones, y no existe un marco de validación estándar (Mori et al., 2015).

En este sentido, la presente investigación tiene como propósito la formulación de una serie de modelos basados en datos históricos y en tiempo real, cuyo propósito será la predicción de los tiempos de llegada de los buses a las estaciones tomando como caso de estudio el sistema troncal de MEGABUS en la ciudad de Pereira. Los enfoques propuestos son el método de velocidad promedio, regresión lineal, redes neuronales artificiales (ANN), las máquinas de vectores de soporte (SVM), la regresión Ridge, la regresión Lasso y además un

método bayesiano propuesto para el desarrollo de esta tesis que considera la actualización de la información histórica (caracterizada por una distribución de probabilidad) a medida que se obtiene nueva información por parte de los GPS de los buses. Cada una de estas metodologías será evaluada en un contexto offline en el cual se cuenta con una base de datos histórica para la estimación del modelo de predicción, y sólo la inferencia bayesiana será evaluada en un contexto online en el cual se cuenta con una base de datos histórica que es nutrida por información en tiempo real. El propósito final de la investigación es realizar una comparación de los resultados obtenidos utilizando cada uno de los métodos y realizar recomendaciones sobre que cual de estos pueden arrojar mejores predicciones considerando además el esfuerzo computacional requerido.

Cabe destacar que hasta el mejor conocimiento de la autora, la regresión Lasso no ha sido empleada para predecir tiempos de llegada o tiempos de viaje en el contexto del tránsito, sino para predecir el tráfico en tiempo real en carretera (Kamarianakis et al., 2012).

3 MODELACIÓN DE LA HORA DE LLEGADA DE LOS BUSES A LOS PARADEROS DEL SISTEMA DE TRANSPORTE BRT

Usando la base de datos de registros GPS históricos de los buses que transitan el sistema troncal brindada por la empresa MEGABUS, a continuación, se presentan modelos para la estimación de la hora de llegada de los buses a los paraderos del sistema de transporte BRT de la ciudad de Pereira. Se incluyen modelos que incorporan un componente histórico y un componente actual.

El objetivo es desarrollar modelos para estimar la hora de llegada de los buses a los paraderos del sistema de transporte de estudio. Primeramente, desde un contexto offline mediante la información histórica disponible, para luego comparar los resultados de dicha predicción de los diferentes modelos, con el método de la inferencia bayesiana, que emplea tanto la información histórica como la del tiempo real de la ubicación del vehículo. Finalmente, se analizarán todos los modelos empleados y se determinará cuál ofrece un mejor desempeño.

En la siguiente sección se hace una descripción de la base de datos de los registros GPS empleados y de las variables estimadas para la estimación de los modelos propuestos. Después se presenta una breve revisión teórica de los modelos escogidos para estimar la hora de llegada de los buses a los paraderos.

Posteriormente se presentan los resultados de los modelos tanto para un contexto pasivo como para el contexto en tiempo real, y así mismo las medidas de desempeño para evaluar el comportamiento de estos.

Finalmente, se hace un análisis de los resultados obtenidos en la estimación de los modelos y se determina cuál debería ser implementado en el sistema.

3.1 Descripción de la base de datos

Para la realización de esta investigación se empleó una base de datos suministrada por la empresa MEGABUS, entidad encargada de la operación del sistema BRT de la ciudad de Pereira, Colombia. Dicha base de datos cuenta con información referente al posicionamiento de los buses del sistema de transporte masivo en el tramo troncal de la red, proporcionada por los dispositivos GPS (por sus siglas en inglés Global Positioning System) incorporados

en los vehículos. Estos envían un registro de la posición del vehículo cada 3 segundos (aproximadamente) al centro de control. Cada registro contiene información acerca del código del bus, la fecha, la hora, la posición del bus (latitud y longitud), el azimut del vehículo y nombre de la ruta que se encuentra realizando.

A continuación, se presenta un ejemplo de la información registrada en la base de datos empleada:

Tabla 1. Ejemplo de información de la base de datos de buses

Nombre	Definición	Ejemplo
REP_GPS_CODIGO	Elemento de identificación del dato	MI051_5299454001367944
EQU_CODIGO	Nombre del bus	MI051
FECHA_GPS	Fecha en la que se registró este dato	17/03/2014
HORA_GPS	Hora a la que se registró este dato	6:21:04
LON_GPS	Componente de la posición geográfica del bus	-75.739243
LAT_GPS	Componente de la posición geográfica del bus	4.8055601
VEL_GPS	Velocidad instantánea del bus	0
DIR_GPS	Azimut del bus	322
ACL_GPS	Aceleración instantánea del bus	0
ODO_GPS	Odómetro del bus	92712672
RUT_NOMBRE	Ruta que se encuentra realizando	Troncal 2

Fuente: Elaboración propia

La información utilizada proviene de dos periodos utilizados con diferentes propósitos: una base de datos con información entre los días 17-30 de marzo de 2014, la cual fue utilizada para la estimación de los modelos, y otra base de datos con información entre los días 1-6 de junio de 2014 utilizada para la validación de los distintos modelos.

Es importante aclarar que, en el caso del modelo de inferencia bayesiana propuesto, se realizó la actualización de la información histórica a medida que se reciben nuevos registros del GPS del bus.

El sistema de Megabus cuenta con tres rutas troncales a lo largo de la ciudad, que fueron incluidas en el objeto de estudio en esta investigación, debido a que eran las únicas que contaban con paraderos específicos para el ascenso y descenso de pasajeros, aspecto fundamental al momento de predecir la hora de llegada de los buses a las estaciones.

A continuación, se presenta el esquema de las rutas troncales y sus respectivas estaciones.

Ilustración 1. Rutas troncales y estaciones de Megabus

Fuente: MEGABUS

La empresa Megabus también cuenta con diferentes rutas periféricas en la ciudad de Pereira, sin embargo, éstas no cuentan con paraderos de ascenso y descenso de pasajeros definidos. Por lo tanto, no fueron incluidas en estas estimaciones.

Es importante hacer claridad que, cuando la información analizada en la presente investigación fue tomada, la operación de este sistema BRT la realizaban dos operadores distintos: Promasivo y Asemtur, y que solamente los buses de la flota de esta última empresa cuentan con dispositivos GPS incorporados en los vehículos, por lo que las estimaciones realizadas tendrán en cuenta sólo la flota en operación de esa empresa.

Una vez revisada la base de datos suministrada, se realizó una depuración de la información para eliminar registros que se encontraran por fuera del área de estudio. Luego de la

depuración, se trabajó con un total de 2.641.205 datos recolectados, los cuales se dividen en 1.801.518 registros del mes de marzo para la estimación de los modelos y 839.687 del mes de junio para la validación de estos. Las Tablas 2 y 3 resumen la información de los registros:

Tabla 2. Información sobre datos GPS del mes de marzo 2014

Fecha	Día de la semana	Número de datos
17/03/2014	Lunes	154379
18/03/2014	Martes	149199
19/03/2014	Miércoles	156282
20/03/2014	Jueves	190083
21/03/2014	Viernes	144016
22/03/2014	Sábado	145071
23/03/2014	Domingo	59412
24/03/2014	Lunes festivo	79150
25/03/2014	Martes	145806
26/03/2014	Miércoles	142408
27/03/2014	Jueves	147294
28/03/2014	Viernes	127405
29/03/2014	Sábado	121185
30/03/2014	Domingo	39828
Total		1801518

Fuente: Elaboración propia

Tabla 3. Información sobre datos GPS del mes de junio 2014

Fecha	Día de la semana	Número de datos
1/06/2014	Domingo	92663
2/06/2014	Lunes festivo	89217
3/06/2014	Martes	159356
4/06/2014	Miércoles	180127
5/06/2014	Jueves	155806
6/06/2014	Viernes	162518
Total		839687

Fuente: Elaboración propia

El proceso de depuración de la información se realizó en el programa ArcGis. Dicho proceso consistió en la conversión de los registros suministrados por el ente operador de tablas de

datos a capas geográficas, para cada uno de los días mencionados, con el fin de desplegarlos en el programa y poder visualizarlos.

Una vez realizado el proceso de conversión, y con el fin de eliminar datos atípicos que puedan generar ruido en el sistema al momento de realizar los modelos de predicción, se establecieron los siguientes criterios para realizar la eliminación de los datos GPS:

1. Los datos registrados de la posición del bus localizado por fuera de geocercas establecidas a lo largo del sistema troncal fueron eliminados. Lo anterior debido a que correspondían al recorrido que hacía el bus desde los patios hacia el intercambiador para comenzar su recorrido o recorridos del bus por fuera del carril exclusivo del sistema masivo, debido a circunstancias operacionales.
2. Los datos registrados de la posición del bus que se encuentren realizando un recorrido dentro del sistema troncal que no se encuentre permitido. Por ejemplo, que un vehículo se encuentre realizando el regreso a uno de los dos intercambiadores, transitando por la secuencia de paraderos errónea.
3. Los datos registrados de la posición del bus no se encuentren describiendo la trayectoria de la ruta que tienen designada en la base de datos. Por ejemplo, en los datos registrados muestra que el bus se encuentra realizando la Troncal 1; sin embargo, al seguir la trayectoria de este se evidencia que operaba sobre la Troncal 2.
4. Los datos registrados de la posición del bus que no contaran con toda la información del recorrido del bus. Por ejemplo, el día 20 de marzo, el bus identificado con el código MI071 tenía programado servicio durante todo el día. Sin embargo, al momento de revisar la información suministrada, se advirtió que los datos no presentaban recorridos continuos y faltaba mucha información, presuntamente por una falla en el GPS del vehículo. Se eliminaron esos datos y no se tuvieron en cuenta para realizar las estimaciones.

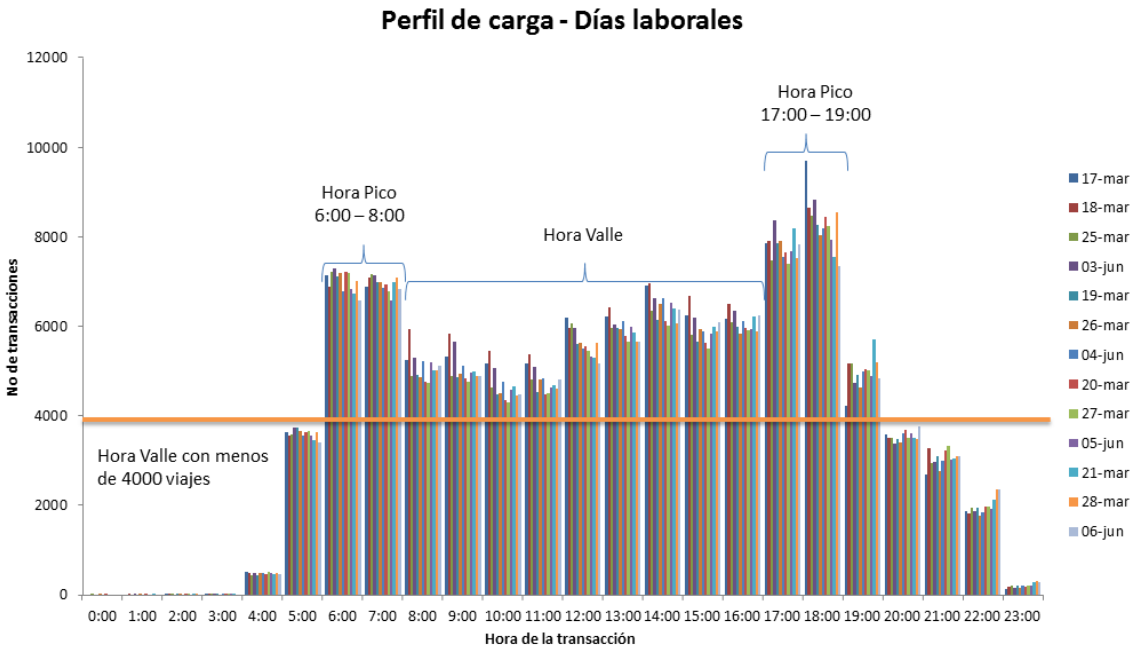
Asimismo, junto con la depuración de datos, se realizó la validación de la información relacionada sobre la programación de la operación de los buses en los días analizados. La programación contenía información sobre el horario en el que el bus se encuentra en servicio, la ruta que se encuentra realizando, las horas estimadas de llegada a los intercambiadores del sistema y el punto de salida del vehículo.

Una vez visualizada la información en el programa, se identificó que se presentaban inconsistencias sobre la programación entregada y el recorrido del bus en la troncal. Por lo anterior, se realizaron algunos ajustes para que hubiese congruencia en los dos archivos, los cuales fueron empleados en los diferentes algoritmos para la estimación de las variables explicativas del modelo.

Debido a la gran cantidad de información que se manejó para determinar las variables explicativas y la calibración de los distintos modelos propuestos en este estudio, se optó por el sistema de administración de bases de datos denominado PostgreSQL, que cuenta con una extensión denominada PostGIS, de gran utilidad para el manejo de datos espaciales, índices espaciales y funciones que operan sobre ellos.

Por otra parte, se empleó la base de datos del registro detallado de las transacciones de pago que realizaron los pasajeros para poder ingresar al sistema de transporte, con el fin de obtener los perfiles de los viajes a lo largo del día en el sistema, y así poder realizar la estimación de los distintos modelos, teniendo en cuenta condiciones operacionales similares.

Ilustración 2. Perfil de carga de días hábiles en el sistema



Fuente: Elaboración propia

Considerando lo anterior, se realizó la segmentación de la información en días laborales y fines de semana, y se agregó una variable muda para evaluar el efecto de los periodos pico

dentro de los distintos modelos. Los periodos pico se presentaron entre las 6:00 AM y las 8:00 AM y las 05:00 PM y las 07:00 PM, coincidiendo con las horas en la que se inician los viajes de ida y de regreso de los usuarios. Esta segregación de información se realizó con el fin de establecer conjuntos de datos con características operacionales similares, debido a que los algoritmos de predicción que se modelaron son muy susceptibles a variaciones en el comportamiento de los datos, causando posibles sesgos en los resultados.

Finalmente, se recolectó información secundaria acerca de la posición geográfica de los paraderos del sistema y de la malla vial de la ciudad, con el fin de estimar las variables de entrada de los modelos de predicción.

3.2 Variables del modelo

A continuación, se explican las diferentes variables empleadas para la estimación de los modelos de predicción del tiempo de llegada de los buses a las estaciones, las cuales fueron definidas a partir de la revisión de la literatura y del criterio del investigador. Adicionalmente se describe el proceso de estimación a partir de la información disponible.

3.2.1 Distancia hasta el siguiente paradero aguas abajo

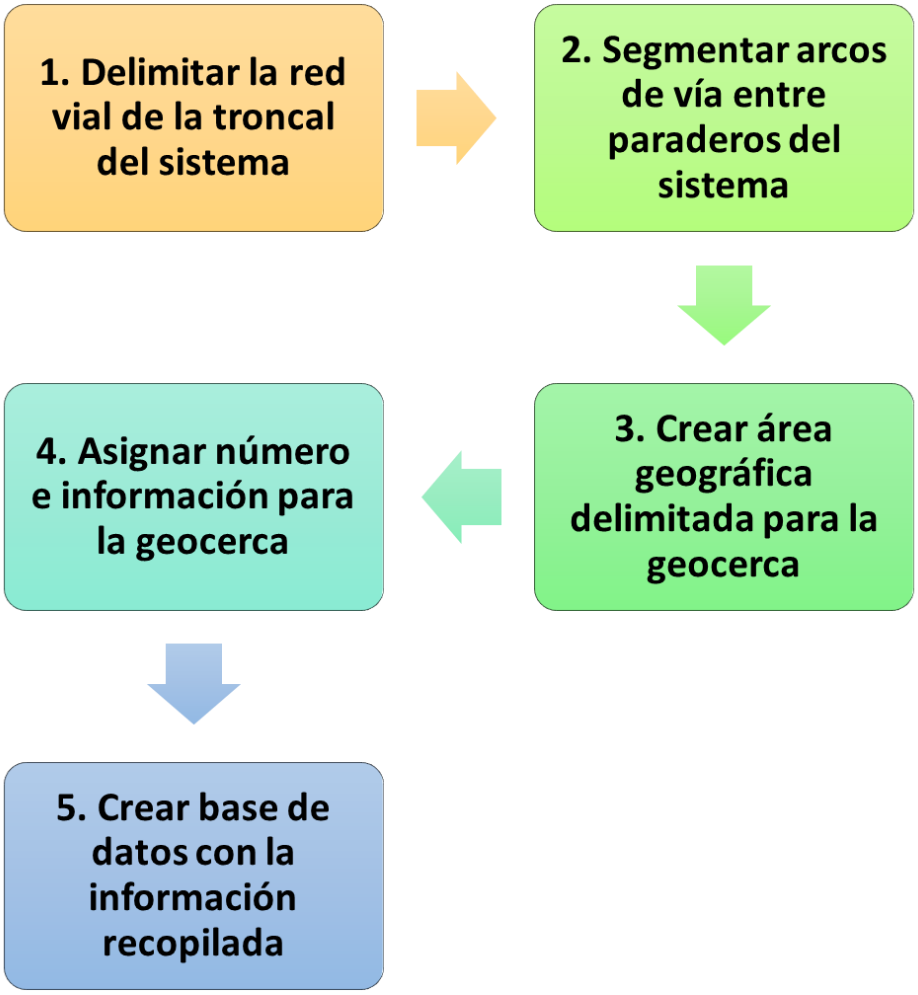
Esta variable es de gran importancia ya que se presenta una relación directamente proporcional con el tiempo de llegada del bus a la estación.

Dado que en la base de datos suministrada tan sólo contenía información sobre las coordenadas geográficas de cada uno de los buses en determinado instante del día, se decidió dividir la malla vial que hace parte del sistema troncal en geocercas, con el fin de obtener información sobre el recorrido de los vehículos por instante del tiempo.

Una geocerca, de acuerdo con la ACM SIGSPATIAL GIS 2013, es un perímetro virtual para una zona geográfica del mundo real que permite a los usuarios recibir notificaciones cada vez que entran o salen de un área especificada (Ravada et al., 2013). Para la construcción de las geocercas se utilizó el programa ArcGis, a partir de la información geográfica obtenida sobre los paraderos del sistema y la red vial de la ciudad.

El procedimiento que se empleó para la creación de las geocercas que delimitarán la red troncal del sistema BRT en estudio, se presenta a continuación:

Ilustración 3. Procedimiento para creación de geocercas



Fuente: Elaboración propia

1. **Delimitar la red vial de la red troncal del sistema:** En este caso, se contaba con la capa geográfica de toda la red vial de la ciudad de Pereira, por lo cual se extrajo la información correspondiente de las vías del sistema troncal.

   Para la selección correcta de la información, se tuvo en cuenta el conocimiento previo que se tenía sobre los recorridos de las rutas troncales. Además, fue de gran ayuda que al visualizar las capas geográficas que corresponden a los datos GPS de los recorridos de los buses, se demarcaban con claridad las vías empleadas por los vehículos.

2. **Segmentar arcos de vía entre los paraderos del sistema:** Luego de analizar el sistema de estudio, de acuerdo con las características del entorno, la longitud total de la red, longitud promedio de paraderos y tiempo promedio en que se recibe un datos nuevos sobre la posición geográfica del bus, se estipula la distancia promedio en la que se segmentarán los tramos de vía. Para este caso de estudio, la distancia establecida fue de 50 metros. Seguidamente, se procede a segmentar los arcos de vía entre dos paraderos consecutivos del sistema para realizar las geocercas cada 50 metros aproximadamente.
3. **Crear área geográfica delimitada para la geocerca:** Una vez se tuviesen los arcos cada 50 metros entre paraderos consecutivos, se utiliza la herramienta denominada *"buffer"*, con la cual se generan un polígono con un radio determinado por cada punto de origen (que en este caso es el arco de vía segmentado), creando así el área geográfica que delimitará la geocerca en ese punto.

   Una vez creado el polígono, se procede encontrar sus puntos extremos que enmarcan el área de la geocerca, para poder establecer las coordenadas geográficas de los mismos.
4. **Asignar número e información para la geocerca:** A cada geocerca creada se asigna un número junto con la información de las variables para describir su lugar dentro del sistema troncal. Para el presente caso de estudio, se tuvo en cuenta las siguientes variables:
   a) Las variables mudas creadas para identificar las rutas habilitadas en dicha geocerca.
   b) Información sobre el nombre y la distancia faltante para llegar al paradero de aguas abajo, para cada uno de los sentidos establecidos.
   c) Los valores de referencia del azimuth para definir el sentido del vehículo. En el caso de estudio se emplean para decidir si se dirige hacia el intercambiador Cuba o Dosquebradas.
   d) La geometría de la geocerca, que consiste en la representación de la geocerca en coordenadas geográficas de acuerdo con el formato empleado por la herramienta de Postgis de la base de datos Postgres.

5. **Crear base de datos con la información recopilada:** Por último, se recopila la información de todas las geocercas del sistema troncal BRT en un tabla que corresponderá a la base de datos de todas las geocercas establecidas.

   A continuación, se presenta un ejemplo de la información registrada en la base de datos constituida para las geocercas del sistema:

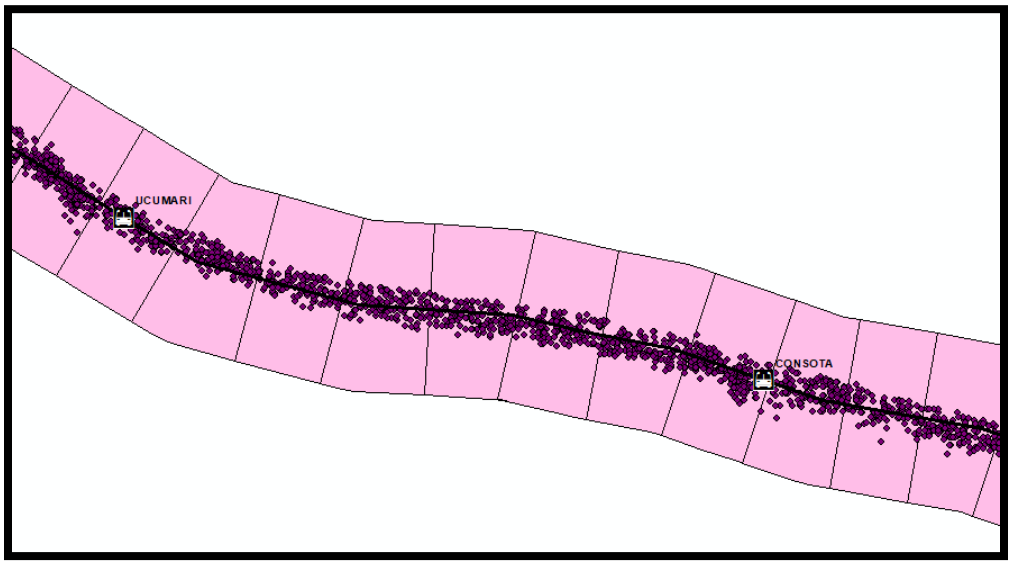
Tabla 4. Ejemplo de información de la base de datos de las geocercas

Nombre	Definición	Ejemplo
id	Número de la geocerca	158
t1	Variable muda. Geocerca habilitada para vehículos que realicen la ruta 1	1
t2	Variable muda. Geocerca habilitada para vehículos que realicen la ruta 2	1
t3	Variable muda. Geocerca habilitada para vehículos que realicen la ruta 3	0
parada1	Nombre de paradero aguas abajo. Recorrido hacia intercambiador Dosquebradas	La Popa
distancia1	Distancia hacia paradero aguas abajo. Recorrido hacia intercambiador Dosquebradas	1150
parada2	Nombre de paradero aguas abajo. Recorrido hacia intercambiador Cuba	Villavicencio
distancia2	Distancia hacia paradero aguas abajo. Recorrido hacia intercambiador Cuba	100
dir1	Rango inferior de azimuth	140
dir2	Rango superior de azimuth	220
geocerca	Coordenadas geográficas que definen la geometría de la geocerca	ST_GeomFromText('POLYGON((4.8134935 -75.6870162799999, 4.813635297 -75.68790624, 4.814158772 -75.6877896, 4.813835398 -75.6869481199999, 4.8134935 -75.6870162799999))')

Fuente: Elaboración propia

En la Ilustración 4 se muestra un ejemplo sobre la representación visual de las geocercas creadas entre las estaciones Ucumari y Consota del sistema en estudio, junto con la capa geográfica de los datos GPS del recorrido del bus.

Ilustración 4. Geocercas entre estaciones Ucumari y Consota



Fuente: Elaboración propia

Una vez establecida toda la base de datos relacionada con las geocercas del sistema, se realizaron diversos algoritmos con el fin de determinar la variable de distancia hasta el siguiente paradero aguas abajo. Estos algoritmos fueron desarrollados en el lenguaje de programación Python. La librería *"psycopg2"* se utilizó para la conexión con las bases de datos desarrolladas en Postgres, lo cual permitió realizar las consultas necesarias y estimar las variables a partir de los registros GPS de los buses para luego ser almacenadas en dichas bases de datos.

A continuación, se presentan el procedimiento realizado para la determinación de dicha variable:

1. El primer algoritmo está programado para que, mediante la herramienta de PostGIS *"ST_Contains"*, se asigne la geocerca en la cual está contenida el registro GPS de la posición geográfica del bus, teniendo en cuenta que la misma se encuentre habilitada para la ruta que está operando el bus (verificación realizada mediante las variables mudas de las troncales).

2. En el segundo algoritmo, una vez asignado el número de la geocerca en la que se encuentra, se determina el sentido hacia donde se dirige el vehículo. Esta información es de gran importancia debido a que, cada geocerca cuenta con la información sobre el próximo paradero aguas abajo, y dependiendo si se dirige hacia el intercambiador Dosquebradas o hacia el intercambiador Cuba, esta información varía. Es importante aclarar que algunas geocercas sólo cuentan con información para uno de los dos sentidos, establecer el sentido correcto hacia donde se dirige el bus es indispensable para determinar si éste se encuentra realizando un recorrido no autorizado.

   Para la asignación de esta variable se tuvieron diversos criterios, los cuales sirvieron como puntos de control, para verificar el sentido de los recorridos de los vehículos en el sistema. A continuación, se presentan los criterios que se tuvieron en cuenta:

   a) Cuando la posición geográfica del bus se detectaba dentro de las geocercas de los intercambiadores de Cuba o Dosquebradas, se realiza el cambio en la variable, dado por entendido que terminaba su recorrido para iniciar uno nuevo.
   b) Las primeras 10 geocercas siguientes a los intercambiadores, cuentan con información sobre valores del rango permitido del azimuth en el que se puede encontrar el valor de la dirección del vehículo para establecer el sentido de trayectoria de éste. Lo anterior con el fin de distinguir los vehículos que llegaban o que salían de los intercambiadores.
   c) En el caso especial de la Troncal 3, debido a que por la trayectoria en el sistema nunca llega al intercambiador de Dosquebradas, se estableció como punto medio el paradero denominado Libertad, donde se realiza el cambio de la variable, dando por entendido que su recorrido terminaba para iniciar uno nuevo.
   d) Para el caso especial de geocercas que sólo cuentan con información para uno de los dos sentidos de circulación, se realizaba la verificación con los valores del rango permitido del azimut en el que se puede encontrar el valor de la dirección del vehículo.
   e) En caso tal de no encontrarse en ninguno de los anteriores criterios, se asignó el sentido del registro inmediatamente anterior del mismo bus.

3. El tercer algoritmo, una vez asignado los valores de la geocerca en la cual está ubicado el vehículo en dicho instante de tiempo y su sentido de circulación, se realiza la consulta en la base de datos general de las geocercas. Allí se verifica que, de acuerdo con el recorrido de la ruta, número de geocerca y sentido de trayectoria previamente asignados, se determine el nombre del paradero siguiente (aguas abajo) y la distancia promedio a éste. En caso tal que la combinación de dichas variables no fuese válida (que para esa ruta no hay asignación para esa geocerca en ese sentido), se realiza una corrección para ajustar la asignada previamente, para que se determine en la geocerca correcta en la que se encuentra en ese instante. En caso de encontrar no encontrar una combinación de variables válida, se procede a realizar la asignación del paradero siguiente (aguas abajo) y la distancia promedio a éste de la geocerca inmediatamente anterior al último registro del mismo bus.

3.2.2 Diferencia de headway

El headway se define como el intervalo de tiempo entre dos buses consecutivos en un paradero determinado. En una situación ideal, el headway entre dos buses consecutivos en cada paradero debe ser constante a lo largo de la misma ruta. Sin embargo, la realidad es que los headways para los diferentes paraderos se vuelven irregulares debido a la congestión vehicular, las demandas inesperadas de los pasajeros, el comportamiento heterogéneo de los conductores de autobuses y los diseños irrazonables de las bahías de autobuses (Yu et al., 2016).

Dado que este tiempo juega un papel importante en la operación de dichos vehículos, se optó por incluirlo como variable explicativa en los modelos de predicción del tiempo de llegada de los buses a las estaciones.

Para este estudio se consideró la diferencia de headway (Δ_h) como la resta entre el headway actual (h_{act}) menos el headway de diseño (h_{dis}) y la sumatoria del tiempo promedio de parada en cada uno de los paraderos que se encuentran entre las dos posiciones (t_{prom_par}).

El headway de diseño (h_{dis}) es el intervalo teórico de tiempo entre dos buses consecutivos de la misma ruta establecido a través de la programación que realiza el operador del sistema.

El headway actual (h_{act}) es el intervalo real de tiempo entre dos buses consecutivos de la misma ruta, que en la situación ideal tendría el mismo valor que la de diseño. Sin embargo, por factores ajenos a la operación del sistema, este se puede ver afectado.

Para realizar el cálculo de ambos headways, se desarrollaron algoritmos en el lenguaje de programación Python. Con la librería *"psycopg2"* se realizó la conexión con las bases de datos desarrolladas en Postgres, para poder realizar las consultas necesarias y estimar las variables a partir de los registros GPS de los buses, para luego ser almacenadas en dichas bases de datos.

Para poder realizar el cálculo de estas variables, es fundamental tomar la información suministrada de la programación brindada por el operador del sistema, y crear dos tablas correspondientes para cada día. En el primer tipo de tabla se toma la información detallada sobre cada uno de los recorridos programados y realizados por cada uno de los buses, junto con la hora de inicio y fin de cada vuelta realizada por el vehículo, el intercambiador por el cual comienza cada vuelta y el nombre de la ruta que se encuentra operando. El segundo tipo de tabla se genera a partir de la información suministrada, dado que se busca contar con una tabla más resumida, donde se muestre la información general sobre los recorridos programados y realizados por cada uno de los buses. Esta tabla contiene la hora de inicio y fin de su operación, el intercambiador por el cual comienza su recorrido, el sentido del recorrido con que inicia su operación y el nombre de la ruta que se encuentra operando

Seguidamente, se procede a establecer el bus inmediatamente anterior mediante un algoritmo que realiza un chequeo con la información consagrada las tablas de programación, de la ruta en la que se encuentra operando el bus, del orden en el que fueron despachados los vehículos y en cuál intercambiador comenzó sus recorridos. Este dato se asigna en la base de datos para cada uno de los registros GPS de la localización del bus.

Una vez establecido el bus inmediatamente anterior al vehículo en análisis, se proceden a calcular los valores actuales y de diseño del headway. Para el caso del valor de diseño, el algoritmo selecciona la información correspondiente a las horas de despacho del vehículo analizado y de su predecesor para realizar la diferencia de estas y establecer el headway. Para el caso del valor actual, el algoritmo el algoritmo selecciona la información correspondiente

a las horas que el vehículo analizado y de su predecesor fueron detectados en la misma geocerca, para realizar la diferencia de estas y establecer el headway

Mientras que, el tiempo promedio de parada en el paradero (t_{prom_par}) hace referencia al tiempo que se demora un bus en una estación para permitir el ascenso y descenso de pasajeros. Se tiene en cuenta las estaciones que se encuentran entre la posición actual de dos buses consecutivos.

Para la asignación de esta variable, fue necesario realizar un algoritmo que determinara el tiempo de parada del bus en cada una de las estaciones que paraba, el cual se calcula como la resta entre las horas en la que se registraron los datos GPS donde por primera vez se estableció que la distancia al paradero era cero y la última antes de continuar su recorrido. Cada uno de estos valores es almacenado en una base de datos, junto con el nombre de la estación correspondiente, el sentido en el que se encuentra circulando el bus y la fecha en la que se registró. Posteriormente, se realiza un análisis de todos los datos recolectados, tanto para los días hábiles como para los días no hábiles, y se establecen los valores promedio para cada una de las estaciones para cada uno de los sentidos evaluados y se crea una base de datos que se consultará para el cálculo de la variable de tiempo promedio de parada en el paradero.

Luego, un algoritmo se encarga de establecer que paraderos se encuentran entre el vehículo analizado y de su predecesor y así sumar los tiempos promedios de parada de éstos.

Así las cosas, se define la diferencia de headway con la siguiente fórmula

$$\Delta_h = h_{act} - h_{dis} - \sum t_{prom\_par}$$

Asimismo, en algunos modelos se estimó esta variable sin la inclusión de la sumatoria del tiempo promedio de parada en el paradero (t_{prom_par}), con el fin comparar si su inclusión resultaba importante en el modelo.

3.2.3 Promedio ponderado del tiempo de llegada de buses anteriores

Esta variable consiste en el promedio ponderado del tiempo de llegada de los buses anteriores de cualquier ruta. El último bus que precede al bus analizado, contribuirá más a la estimación del tiempo que los anteriores a él (Yu et al., 2011).

En este caso particular, se propuso una ponderación inversa respecto al tiempo transcurrido entre los últimos tres buses anteriores al bus analizado. A continuación, se presenta la fórmula para realizar la estimación de la variable:

$$t_{pond} = \sum_{i=1}^{3} t_{ik} \times \left(\frac{T_{total} - T_{ik}}{T_{total}}\right)$$

Donde t_{ik} es el tiempo que se tardó el bus i en llegar a la siguiente estación estando en la geocerca k, T_{total} es la sumatoria de los tiempos en los que se registró que pasaron los últimos tres buses que pasaron por la geocerca k y T_{ik} es el tiempo en el que el bus i en pasó por la geocerca k.

Con este planteamiento, se busca obtener información sobre las condiciones operacionales del sistema, para generar una mejor y más acertada predicción.

El algoritmo programado para el cálculo de esta variable toma información de una base de datos generada con información sobre el promedio histórico del tiempo de llegada del bus a la próxima estación cuando se encontrara en la geocerca k. A medida del transcurso de la operación de los buses a lo largo del día, se va recopilando y actualizando la información de estos tiempos, para realizar los cálculos necesarios para la determinación de este valor.

3.2.4 Periodo pico

Es una variable muda, la cual tiene como objetivo evaluar el efecto que tiene la hora pico al momento de incluirla en los modelos de predicción.

La información para determinar el periodo pico del sistema mediante el registro detallado de las transacciones de pago que realizaron los pasajeros, como se había mencionado anteriormente.

Los periodos pico identificados se presentaron entre las 6:00 AM y las 8:00 AM y las 05:00 PM y las 07:00 PM, coincidiendo con las horas en la que se inician los viajes de ida y de regreso de los usuarios, por lo que en la base de datos de la información para los registros GPS de las coordenadas geográficas tiene una columna donde se asigna 1 si el dato está dentro de esos periodos, de lo contrario se asigna 0.

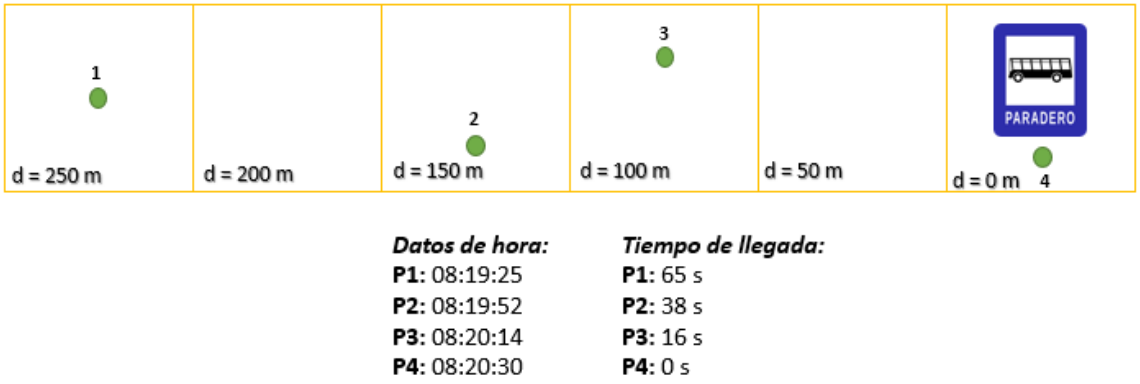
3.2.5 Tiempo de llegada del bus

Esta variable es crucial para realizar los algoritmos de predicción, debido a que, como tan sólo se cuenta con la información de los registros GPS de los buses, no se tiene con antelación la variable que se busca estudiar y predecir.

El tiempo de llegada de los buses a las estaciones se calculó a partir del registro de la coordenada geográfica enviada por el GPS junto con la estimación realizada de la variable de distancia hasta el siguiente paradero aguas abajo.

El algoritmo desarrollado para el cálculo de esta variable consistió en identificar la hora de aquellos registros donde la distancia al paradero fuese cero, es decir que llegó a la estación, y tomarlo como punto de referencia para realizar la diferencia con la hora del dato de registro en la posición donde aún no ha llegado, como se muestra en la Ilustración 5.

Ilustración 5. Tiempo de llegada del bus



Fuente: Elaboración propia

Es importante destacar que, con el fin de corroborar la eficiencia de los modelos que sólo proponen predecir el tiempo de llegada con las variaciones que éste puede tener y comprobar el ajuste de los métodos, se desarrollaron modelos de predicción donde se empleó exclusivamente esta variable.

Para estos modelos, la información fue segmentada para cada una de las rutas del sistema, debido a que éstas realizan un recorrido diferente, por lo que el número de geocercas en la cual pueden transitar es diferente. La información fue analizada en matrices, en las cuales cada fila correspondía a un recorrido del bus, y las columnas a cada geocerca.

Asimismo, se realizó el análisis de los modelos diferenciando los días hábiles y no hábiles de la información disponible.

Dado que existe la posibilidad que a dos registros diferentes del GPS se les asigne la misma geocerca, el algoritmo para la creación de las matrices tan sólo tendrá en cuenta el primer dato. También asignará una nueva fila, indicando el inicio de un nuevo recorrido, una vez llegue a la geocerca que corresponde al intercambiador final.

3.3 Modelos de predicción

Basándonos en la revisión realizada en el estado del arte sobre la predicción del tiempo de llegada de los buses a las estaciones, y de analizar la información disponible, se definieron dos etapas para realizar la estimación:

i. Se desarrollaron distintos modelos de predicción únicamente utilizando la información histórica, en un contexto pasivo (offline). Los modelos propuestos para esta primera parte son velocidad promedio, regresión lineal, red neuronal artificial (ANN), máquina de vectores de soporte (SVM), k vecinos más próximo (kNN), regresión LASSO y regresión ridge.
ii. Se buscó implementar una metodología de inferencia bayesiana que no se haya empleado; la práctica común es emplear filtros de Kalman. A partir de la función de probabilidad de la información histórica disponible, se obtiene la predicción del tiempo de llegada. Para hacerlo, se formula un modelo dinámico que actualiza la

información histórica considerando la información que está siendo observada en tiempo real.

Los distintos modelos de predicción fueron desarrollados usando el lenguaje de programación Python. A continuación, se presenta una descripción de los modelos propuestos:

3.3.1 Velocidad promedio:

Este tipo de modelos utilizan la velocidad media de los vehículos en determinados tramos viales para predecir los tiempos de viaje. Se aplican principalmente para predecir tiempo de viaje cuando se tiene disponibilidad de datos recopilados por tecnología GPS, ya que la distancia recorrida se puede calcular utilizando la información de la posición geográfica (Altinkaya & Zontul, 2013).

A continuación, se presenta la ecuación de velocidad empleada para esta metodología:

$$V_{ij} = \frac{\Delta d_{ij}}{\Delta t_{ij}}$$

Donde:

La posición i del bus hace referencia al momento en el que retoma su recorrido hacia la siguiente estación aguas abajo

La posición j del bus hace referencia al momento en que llega a la estación aguas abajo

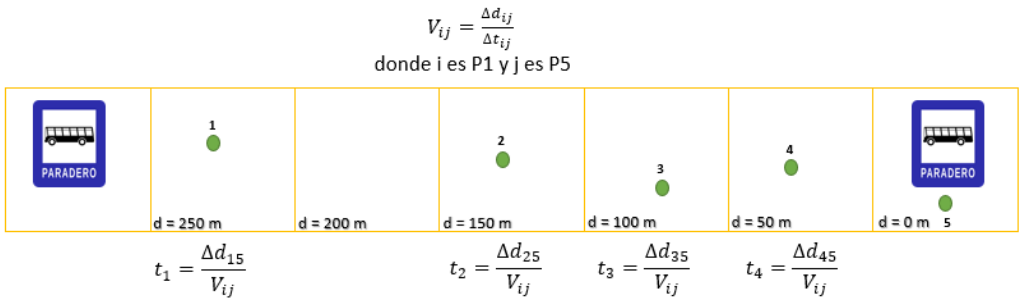
V_{ij} es la velocidad promedio de viaje entre dos paraderos consecutivos entre las posiciones i y j.

Δd_{ij} es la diferencia de distancia entre las posiciones i y j.

Δt_{ij} es la diferencia de horas entre las posiciones i y j.

Para emplear el cálculo de la velocidad promedio en la estimación del tiempo de llegada de los buses a las estaciones, se desarrolló un algoritmo que estimara la velocidad promedio en los tramos viales de dos estaciones consecutivas, y aplicar este valor junto con la distancia hasta el siguiente paradero aguas abajo para la estimación del tiempo.

Ilustración 6. Velocidad promedio



Fuente: Elaboración propia

3.3.2 Regresión lineal:

La regresión lineal hace parte de la familia de los modelos estadísticos, los cuales están enfocados en determinar las variables que inciden en la predicción del tiempo tales como la distancia, el tiempo de llegada de buses anteriores, el tiempo de parada, el comportamiento del conductor, las intersecciones, las señales de tránsito, entre otras.

La precisión de estos métodos depende de que todas las variables que expliquen el modelo sean identificadas e incorporadas en el mismo, lo que resulta ser un procedimiento difícil. Sin embargo, una de las ventajas de la regresión multilineal es que revela qué variables de entrada son menos o más importantes para la predicción (Altinkaya & Zontul, 2013).

La regresión lineal es el primer tipo de análisis de regresión y se usa ampliamente en aplicaciones prácticas, el cual consiste en modelar la relación entre el tiempo de llegada del bus a la estación (variable dependiente) y los factores de impacto (variables independientes) (Yu et al., 2011).

A continuación, se presenta la ecuación de tipo empleada para esta metodología:

$$t_i = \beta_1 \times V_1 + \beta_2 \times V_2 + (\dots) + \beta_n \times V_n + \varepsilon$$

Donde:

t_i es el tiempo de llegada el bus a la estación, que en este caso sería la variable dependiente.

V son las variables independientes que se emplearán para explicar el comportamiento de t_i.

β son los coeficientes estimados por la regresión para cada una de las V.

ε es un término aleatorio de la regresión.

Para la estimación de esta metodología, se desarrolló un algoritmo, que junto con la librería sklearn de Python, permitió realizar las diversas estimaciones de modelos de regresión lineal a partir de las variables descritas en la sección 3.2 del presente documento.

Es importante destacar que la aplicabilidad de los modelos de regresión es limitada, debido a que las variables empleadas en los sistemas de transporte están altamente correlacionadas entre ellas (Chien et al., 2002).

3.3.3 Redes neuronales artificiales

La red neuronal artificial (ANN por sus siglas en inglés) es un modelo de aprendizaje de máquinas inspirado en la capacidad de procesamiento de datos de la estructura neuronal del cerebro humano. Está construido a partir de múltiples capas de neuronas artificiales, que tienen funciones de activación asociadas con sus pesos de entrada. La información puede ser procesada hacia delante o hacia atrás a través de elementos de computación neuronal parcial o totalmente conectados (Ma et al., 2019).

La arquitectura de la ANN está generalmente compuesta por un conjunto de nodos y conexiones organizados en capas. Para los modelos estimados en el presente estudio, se utilizaron tres capas: la capa de entrada, oculta y de salida.

La capa de entrada se utiliza para recibir el conjunto de datos de entrenamiento en la red, que en este caso son las variables explicativas del modelo. Mientras que la capa de salida es una sola neurona que en la cual obtiene la predicción del tiempo de llegada del bus.

La formulación de la metodología es:

$$Y_j = \psi_j \left(\sum_{i=1}^{m} w_{ji} X_i + b_j \right)$$

Donde:

m es el número de entradas aplicadas a la neurona j.

X_i es el conjunto de variables de entrada de la neurona j.

Y_j es la salida de la neurona j.

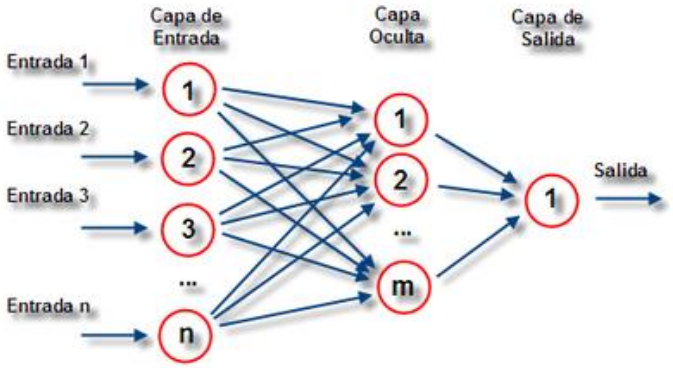
w_{ji} es el peso sináptico que conecta la entrada i a la neurona j.

b_j es el término de error.

$\psi_j(\cdot)$ es la función de activación.

La estructura del modelo se presenta en la Ilustración 7.

Ilustración 7. Estructura propuesta por el modelo RNA



Fuente: Elaboración propia

Para la estimación de esta metodología, se desarrolló un algoritmo, que junto con la librería sklearn de Python, permitió realizar las diversas estimaciones de modelos de ANN a partir de las variables descritas en la sección 3.2 del presente documento.

En este estudio se presenta una red de retroalimentación de dos capas completamente conectada. La función de activación escogida fue la tangente hiperbólica, y se utilizó un algoritmo para resolver los pesos optimizados llamado “adam”, el cual es un optimizador estocástico basado en gradiente propuesto por Kingma, Diederik y Ba.

Una de las ventajas de este modelo es que tiene la capacidad de capturar relaciones complejas entre variables que pueden surgir dentro de grandes cantidades de datos, procesar relaciones no lineales entre predictores y procesar datos complejos y ruidosos (Gurmu & Fan, 2014).

Asimismo, otra ventaja que ofrecen las ANN es que se pueden estimarse sin especificar la forma de la función, mientras que las restricciones sobre la multicolinealidad de las variables explicativas pueden ser ignoradas (Altinkaya & Zontul, 2013).

3.3.4 Máquina de vectores de soporte

La máquina de vectores de soporte (SVM por sus siglas en inglés) son un conjunto de métodos de aprendizaje supervisados relacionados que se utilizan para la clasificación y la regresión (Altinkaya & Zontul, 2013).

Es un algoritmo de aprendizaje de máquina que implícitamente asigna los datos de entrenamiento de un espacio vectorial de entrada de baja dimensión a un espacio de características de mayor dimensión (Ma et al., 2019).

Cuando los puntos de datos no se pueden separar linealmente, SVM los asigna a un espacio de mayor dimensión para que haya un límite de hiperplano entre ellos. En este estudio se empleó el método de regresión de vectores de soporte (SVR por sus siglas en inglés) que es una versión del SVM, utilizando este método para estimar una regresión, en lugar de una clasificación (Philip et al., 2018).

En estudios previos realizados (Vanajakshi & Rilett, 2007; Yu et al., 2011), se ha demostrado que la función kernel RBF proporciona un mejor rendimiento en la predicción del tiempo de viaje, por lo cual fue empleada en el presente estudio.

Para la estimación de esta metodología, se desarrolló un algoritmo, que junto con la librería sklearn de Python, permitió realizar las diversas estimaciones de modelos de SVR a partir de las variables descritas en la sección 3.2 del presente documento. En este estudio se utilizó la función de kernel RBF.

La formulación de la metodología es:

$$K(x_i, x_j) = e^{(-\gamma \|x_i - x_j\|^2)}, \gamma > 0$$

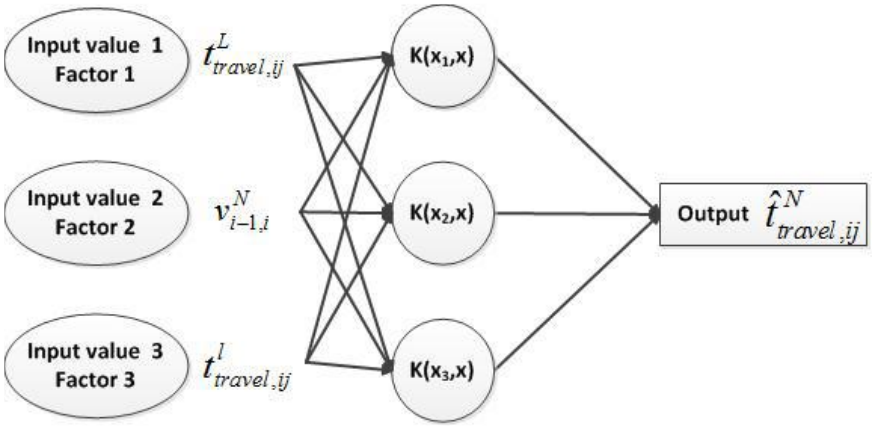
Donde:

γ es un parámetro que debe determinarse de acuerdo con sus necesidades de predicción.

x_i, x_j son los valores de entrada de la máquina de vectores de soporte.

La estructura del modelo se presenta en la Ilustración 8.

Ilustración 8. Estructura propuesta por el modelo SVR



Fuente: A prediction model of bus arrival time at stops with multiple-routes

SVM y SVR han demostrado obtener mejores resultados que el análisis de series de tiempo y el aprendizaje estadístico (Wu et al., 2004). Sin embargo, una de las desventajas de este tipo modelos es que cuando son utilizados para resolver problemas de gran tamaño, se requerirá una gran cantidad de tiempo de cálculo (Bin et al., 2006).

3.3.5 K vecinos más próximos

El método de k vecinos más próximos (kNN por sus siglas en inglés) es uno de los métodos más simples de reconocimiento de patrones. En primera instancia, mide la distancia entre los datos conjunto de prueba y todos los datos del conjunto de entrenamiento. Luego elige los *k* datos más cercanos a los datos conjunto de prueba, arrojando como resultado del modelo la media ponderada de los resultados para todos esos *k* datos del conjunto de entrenamiento (Xin & Chen, 2016).

El kNN es un método no paramétrico utilizado para clasificación o regresión. Para encontrar el vecino más cercano y asignar peso a las contribuciones de los vecinos es útil implementar el algoritmo de distorsión de tiempo dinámico (DTW sus siglas en inglés), el cual calcula la distancia euclidiana entre el vector de registros de la variable de salida y las muestras. La teoría es que el vecino más cercano contribuye más que el más distante a la salida. Por ejemplo, un esquema de ponderación le da a cada vecino un peso de 1/d, donde d es la distancia al vecino (Ma et al., 2019).

La formulación de la metodología es:

$$t = \sum_{j=1}^{k} \frac{1/d_j}{D} (t_{pond})$$

$$d_j = \sqrt{\frac{\lambda_1 \times (\Delta V_1)^2 + \cdots + \lambda_n \times (\Delta V_n)^2}{\lambda_1 + \cdots + \lambda_n}}$$

$$D = \sum_{j=1}^{k} \frac{1}{d_j}$$

Donde:

d_j representa la distancia ponderada entre el dato en la posición j del vecino más cercano y el valor de entrada.

D representa la suma de la distancia ponderada de los k vecinos más cercanos.

λ_n representan los pesos de las variables.

Para la estimación de esta metodología, se desarrolló un algoritmo, que junto con la librería sklearn de Python, permitió realizar las diversas estimaciones de modelos de kNN a partir de la variable del tiempo de llegada del bus a la estación descrita en la sección 3.2.5 del presente documento. En este estudio se utilizaron distintos *k* para obtener el mejor desempeño del modelo, el cual será indicado junto con los resultados del modelo.

3.3.6 Regresión Lasso

LASSO es una versión limitada de los métodos de estimación ordinarios y, al mismo tiempo, un procedimiento automático de construcción de modelos ampliamente utilizado (Kamarianakis et al., 2012). En comparación con los métodos clásicos de selección de variables, como la selección de subconjuntos, LASSO tiene dos ventajas. Primero, el procedimiento de selección es continuo y por lo tanto más estable que la selección de subconjuntos la cual es discreta (Zhao & Yu, 2006). Segundo, LASSO es factible computacionalmente para datos de alta dimensión. En contraste, el cálculo en la selección de subconjuntos es combinatorio y no es factible cuando el número de predictores es muy grande (Madigan, 2008).

LASSO es un método de contracción como ridge, con diferencias sutiles pero importantes. Este método traduce cada coeficiente por un factor constante λ, truncando en cero. Esto se denomina "umbral suave" y se utiliza en el contexto del suavizado basado en ondículas (Hastie, 2009).

La formulación de la metodología es:

$$Y_n = X_n\beta^n + \varepsilon_n$$

$$\hat{\beta}^n(\lambda) = argmin_\beta \|Y_n - X_n\beta\|_2^2 + \lambda\|\beta\|_1$$

Donde:

ε_n es un vector de variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza σ^2

Y_n es la variable de salida

X_n son las variables de entrada del modelo

β^n son los coeficientes del modelo

$\|.\|_1$ representa la norma L1 de un vector que es igual a la suma de los valores absolutos de las entradas del vector.

El parámetro $\lambda \geq 0$ controla la cantidad de regularización aplicada a la estimación.

Para la estimación de esta metodología, se desarrolló un algoritmo, que junto con la librería sklearn de Python, permitió realizar las diversas estimaciones de modelos de regresión LASSO a partir de la variable del tiempo de llegada del bus a la estación descrita en la sección 3.2.5 del presente documento.

3.3.7 Regresión Rigde

La regresión Rigde reduce los coeficientes de regresión al imponer una penalización a su tamaño. Los coeficientes de Rigde minimizan una suma de cuadrados residual penalizada (Hastie, 2009).

La regresión de cresta resuelve los problemas de optimización relacionados con mínimos cuadrados (en modelos de regresión lineal o regresión logística) y máxima verosimilitud (en modelos de elección discreta) al incorporar una restricción cuadrática en los parámetros que se estiman. Esta restricción implica que el nuevo estimador será más eficiente, con menor MSE, pero con un sesgo a la baja en valor absoluto. Sin embargo, dado cómo se ha estimado tradicionalmente el VOT (como la tasa marginal de sustitución entre el tiempo de viaje y el costo del viaje), la restricción agregada al problema de optimización asociado con la estimación de parámetros individuales puede generar estimadores insesgados del VOT, superando el clásico problema de sesgo, de regresión de la cresta mientras que al mismo tiempo mantiene sus propiedades de mayor eficiencia (de Grange et al., 2015).

La formulación de la metodología es:

$$RSS_{ridge} = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde:

λ es un parámetro que controla el grado de penalización: cuanto mayor éste, los coeficientes serán menores resultando más robustos a la colinealidad.

β_j son los coeficientes del modelo.

y_i es la variable de salida.

$f(x_i)$ son las variables de entrada del modelo

Para la estimación de esta metodología, se desarrolló un algoritmo, que junto con la librería sklearn de Python, permitió realizar las diversas estimaciones de modelos de regresión Ridge a partir de la variable del tiempo de llegada del bus a la estación descrita en la sección 3.2.5 del presente documento.

3.3.8 Inferencia bayesiana

La red bayesiana es un gráfico acíclico dirigido que consta de nodos y bordes dirigidos, donde los nodos representan la variable de estado y los bordes dirigidos representan las

dependencias, y la probabilidad condicional entre los nodos padre e hijo determina la fuerza de la asociación entre nodos (Deng et al., 2013).

Para la aplicación de la inferencia bayesiana, se propuso modelar los datos históricos del tiempo de llegada de los buses a las estaciones por medio de una distribución de probabilidad. Luego de estudiar la naturaleza de los datos y realizar las gráficas correspondientes, se escogió una distribución normal, la cual se caracteriza por la media μ y la matriz covarianza B, tal y como se presentan en las siguientes ecuaciones:

Función del prior (datos históricos del tiempo de llegada de los buses a las estaciones)

$$f(\cdot) = N(\mu, B)$$

Probabilidad de la estimación

$$P(x_i|x_{i-1}) \propto P(x_i) \cdot \mathcal{L}\big((x_i|x_{i-1})\big)$$

Formulación del prior

$$P(x_i) \propto e^{-\frac{1}{2}\|x_i - \bar{x}_i^b\|_{B^{-1}}}$$

Función de verosimilitud (likelihood)

$$\mathcal{L}\big((x_i|x_{i-1})\big) \propto e^{-\frac{1}{2}\|x_{i-1} - Hx_i\|_{R^{-1}}}$$

Considerando las funciones previamente especificadas, se define la probabilidad de la estimación con la siguiente ecuación:

$$P(x_i|x_{i-1}) \propto P(x_i) \cdot \mathcal{L}\big((x_i|x_{i-1})\big)$$

$$P(x_i|x_{i-1}) \propto e^{-\frac{1}{2}\|x_i - \bar{x}_i^b\|_{B^{-1}} - \frac{1}{2}\|x_{i-1} - Hx_i\|_{R^{-1}}}$$

$$\bar{x}_i^a = [B^{-1} + H^T R^{-1} H]^{-1}[B^{-1}\bar{x}^b + H^T R^{-1} x_{i-1}]$$

Donde:

$\bar{x}_i^a$ es el valor asimilado.

B es la matriz de covarianza de la información histórica.

H es la matriz observacional.

R es la matriz de covarianza de las observaciones.

$\bar{x}^b$ es la media de la información histórica.

x_{i-1} es la información histórica.

El planteamiento anterior tiene como objetivo no replicar el uso tradicional del bayesiano. En particular, se modela el prior con funciones de densidad de probabilidad, y se asume que el error de las observaciones temporales se trabaja con una distribución normal también. Lo que ocurre inmediatamente es que la distribución posterior (es decir, el dato asimilado) también corresponderá a una distribución normal, donde existen formas cerradas para estimar los momentos posteriores, obteniendo un mejor desempeño.

De igual forma, es importante destacar que, para el planteamiento de este método, se le dio más peso a la observación, y se dedujo que había que ajustar el R en 0.01, indicado por el nivel de confianza para no destruir la información estadística presente. Así las cosas, no sólo se predice el tiempo, sino que también se provee un margen de incertidumbre alrededor de ese tiempo. En el caso de considerar que la predicción realizada presenta un margen de error 0, se destruiría toda la información estadística presente.

Asimismo, se hace claridad que los parámetros del modelo se ajustan a medida que los instantes de tiempo aumentan en el modelo. Es decir, a medida que el bus va desplazándose a través de diferentes geocercas, esta información es incluida en el prior haciendo que la incertidumbre decrezca.

3.4 Estructura de los modelos

Una vez establecidas las diferentes variables explicativas identificadas y las distintas metodologías para realizar la predicción del tiempo de llegada de los buses a las estaciones, en esta sección se presentan las distintas combinaciones empleadas para realizar una comparación entre los resultados obtenidos, y así sacar conclusiones e inferir información valiosa a partir de los parámetros que brinde cada modelo.

Tabla 5. Combinación de métodos y variables explicativas

		VARIABLES				
		Distancia hasta el siguiente paradero aguas abajo	Diferencia de headway	Promedio ponderado del tiempo de llegada de buses anteriores	Hora pico	Tiempo de llegada del bus
MÉTODOS	Velocidad promedio	x				
	Regresión lineal	x	x	x	x	
	ANN	x	x	x	x	
	SVM	x	x	x	x	
	kNN					x
	Regresión Lasso					x
	Regresión Ridge					x
	Inferencia bayesiana					x

Fuente: Elaboración propia

Como se puede observar en la Tabla 5, se presentan cuáles fueron las variables explicativas empleadas para cada uno de los métodos escogidos para este estudio. Se observa que para el método de la velocidad promedio, se utiliza únicamente la variable de distancia para la estimación.

Asimismo, para los métodos como la regresión lineal, ANN y SVM se realizaron modelos empleando las cuatro variables explicativas identificadas. Y a su vez, se realizaron dos (2) tipos de modelos adicionales segmentado por: días hábiles y días no hábiles.

Para los métodos de kNN, Lasso, regresión Ridge y la inferencia bayesiana se empleó únicamente la información histórica del tiempo de llegada de los buses a las estaciones y a cada una de las geocercas del trayecto, con el fin de comprobar lo enunciado en el estado del arte, que se puedan plantear modelos eficientes que ignoren la relación entre el tiempo de viaje y los factores de influencia que afectan su variabilidad.

A continuación, en la Tabla 6 se muestran las diferentes combinaciones empleadas en los modelos para realizar la estimación.

Tabla 6. Modelos estimados para las variables explicativas

Modelo	**Variables de entrada**				
	Distancia hasta el siguiente paradero aguas abajo	Diferencia de headway	Diferencia de headway (sin tiempos promedio de parada)	Promedio ponderado del tiempo de llegada de buses anteriores	Hora pico
1	x		x	x	x
2	x		x	x	
3	x			x	
4				x	
5	x	x		x	x
6	x	x		x	

Fuente: Elaboración propia

La razón por la cual se realizaron las diferentes combinaciones de variables explicativas para este tipo de modelos es que se busca inferir a través de los resultados obtenidos, cuáles son las variables más significativas a la hora de predecir dicho tiempo.

En cuanto a los modelos que sólo emplearon información histórica de los tiempos de llegada, se realizó la segmentación de los datos de acuerdo con la ruta troncal que hiciera el vehículo: Troncal 1, Troncal 2 y Troncal 3. Y a su vez, se realizaron tres (3) tipos de modelos adicionales segmentado por: día completo, días hábiles y días no hábiles.

3.5 Contextos evaluados

Una vez definida la combinación de métodos y variables explicativas y los modelos estimados para las variables explicativas, es importante entender la diferencia entre los dos contextos propuestos para la evaluación de las predicciones.

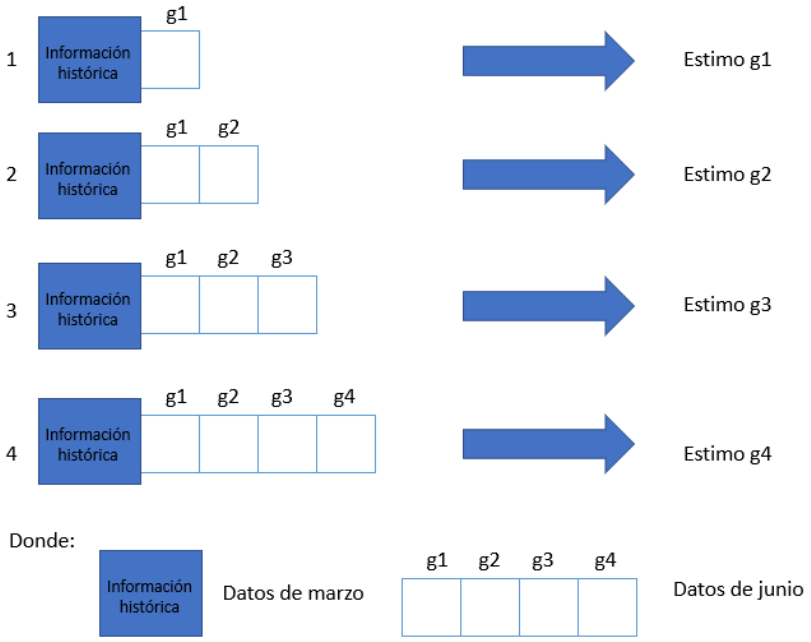
A continuación, se presenta la explicación sobre cómo funciona la estimación y la validación para cada uno de los contextos propuestos en la presente investigación.

3.5.1 Contexto offline

En este contexto, se utiliza la información histórica para estimar los modelos de predicción junto con sus respectivos parámetros, y luego con nuevos registros GPS de los buses se valida la capacidad predictiva de los modelos.

Para los modelos de inferencia bayesiana en este contexto, se obtienen registros GPS de los buses en tiempo real (a medida que el bus se va moviendo), los cuales son empleados para realizar la predicción del tiempo de llegada al punto siguiente con la base de datos histórica que se usa como prior.

Ilustración 9. Proceso de validación para contexto offline para el método de inferencia bayesiana



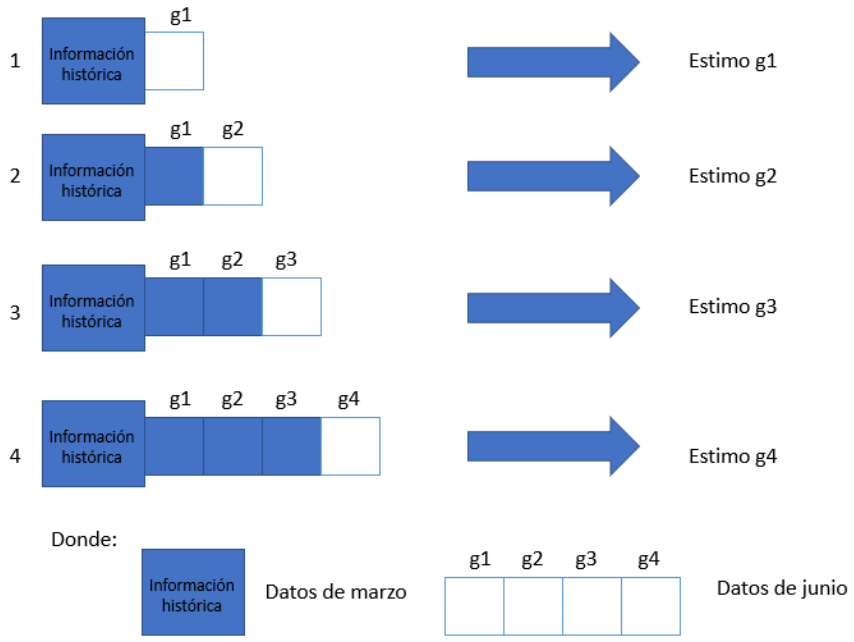
Fuente: Elaboración propia

Para la estimación en el contexto offline, la información del mes de marzo fue empleada como la *prior* inicial del modelo, mientras que la información del mes de junio fue utilizada para validar con la información en tiempo real recibida.

3.5.2 Contexto online

En este caso se obtienen registros GPS de los buses en tiempo real (a medida que el bus se va moviendo); cada dato nuevo registrado es incorporado a la base de datos histórica que se usa como prior para la predicción del tiempo de llegada al punto siguiente. Dado que el bus eventualmente llegará al punto para el cual se acaba de realizar la predicción, es posible realizar una validación entre lo que se predijo y la información real medida por el GPS. Este tipo de metodología online es provechosa porque se alimenta tanto de la distribución de los datos históricos como de lo que el mismo bus viene reflejando en su recorrido actual.

Ilustración 10. Proceso de validación para contexto online para el método de inferencia bayesiana



Fuente: Elaboración propia

Para la estimación en el contexto online, la información del mes de marzo fue empleada como la *prior* inicial del modelo, mientras que la información del mes de junio fue utilizada para validar y alimentar la *prior* con la información en tiempo real recibida. Esto lo convierte en un modelo dinámico, que no sólo emplea la información histórica, sino que tiene en cuenta las condiciones actuales de la red por la nueva información incorporada.

3.6 Medidas de desempeño

Los resultados de la predicción del tiempo de llegada de los distintos modelos anteriormente propuestos se evalúan en términos del rendimiento de tres medidas: el error absoluto medio (EAM), el error porcentual absoluto medio (EPAM) y la raíz del error cuadrático medio (RECM). Los tres términos comparan la diferencia entre el tiempo de observado y el tiempo estimado.

A continuación, se presentan las ecuaciones que definen a cada una de los estimadores:

- *Error absoluto medio (EAM)*

$$EAM = \frac{\sum|t_{observado} - t_{estimado}|}{N}$$

- *Error porcentual absoluto medio (EPAM)*

$$EPAM = \frac{1}{N}\sum \frac{|t_{observado} - t_{estimado}|}{t_{observado}}$$

- *Raíz del error cuadrático medio (RECM)*

$$RECM = \sqrt{\frac{\sum(t_{observado} - t_{estimado})^2}{N-1}}$$

Donde:

$t_{observado}$ es el tiempo de llegada real observado del bus a la estación.

$t_{estimado}$ es el resultado de la predicción del tiempo de llegada del bus a la estación.

N es el número total de datos que han sido observados.

4 RESULTADOS DE LOS MODELOS

En la presente sección se presentarán los resultados obtenidos a partir de los diferentes modelos y métodos seleccionados, con el fin de realizar las comparaciones correspondientes y analizar los hallazgos del estudio.

Primeramente, con el fin de conocer el grado de asociación lineal entre las variables explicativas seleccionadas, y entre estas y la variable dependiente (tiempo de llegada de los buses a las estaciones), se calcularon los coeficientes de correlación, los cuales se presentan a continuación:

Tabla 7. Coeficientes de correlación entre variables explicativas

	Tiempo de llegada	***Hora Pico***	***Distancia***	***Tiempo de llegada ponderado***	***Headway 2***	***Headway 1***
Tiempo de llegada	1					
Hora Pico	0.018	1				
Distancia	0.740	-0.001	1			
Tiempo de llegada ponderado	0.817	0.013	0.786	1		
Headway 2	0.027	-0.045	0.019	0.030	1	
Headway 1	0.025	-0.038	0.018	0.028	0.995	1

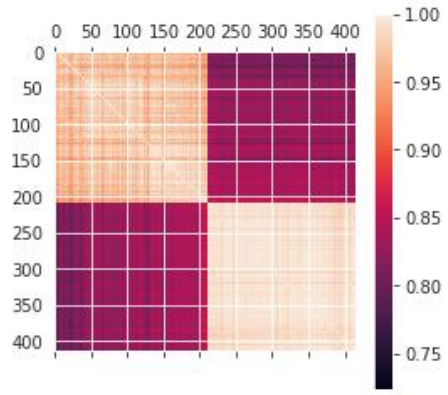
Fuente: Elaboración propia

Los resultados obtenidos indican que, para el contexto evaluado, las variables mayormente correlacionadas con el tiempo de llegada de los buses a las estaciones son la distancia hasta el paradero aguas abajo y el promedio ponderado del tiempo de llegada de los buses anteriores. Otro aspecto importante que se puede observar de los coeficientes de correlación es que la variable muda que captura los periodos pico del sistema resultó tener muy poca incidencia en los modelos, por lo cual se optó por estimar modelos distinguiendo únicamente entre los días hábiles y no hábiles.

De igual forma, se realizó el análisis de la correlación existente entre las horas de llegada de los buses a las distintas geocercas establecidas, con el fin de identificar el grado de asociación espacial de los tiempos de llegada en cada recorrido.

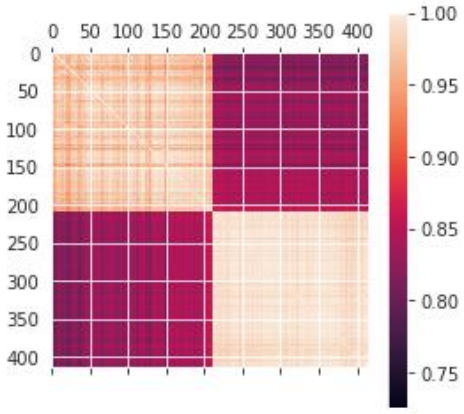
A continuación, se presentan las gráficas de correlación espacial:

Ilustración 11. Gráfica de coeficientes de correlación para todos los datos de Troncal 1



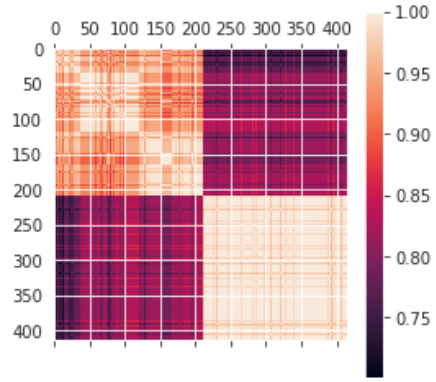
Fuente: Elaboración propia

Ilustración 12. Gráfica de coeficientes de correlación para datos de días hábiles de Troncal 1



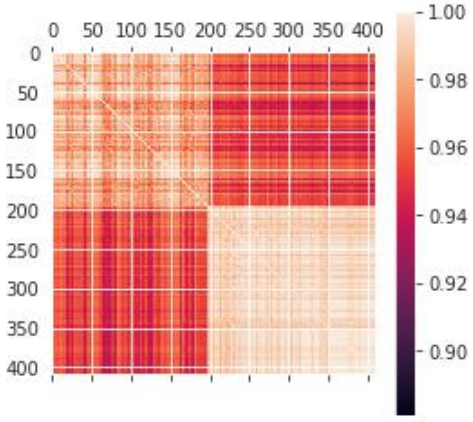
Fuente: Elaboración propia

Ilustración 13. Gráfica de coeficientes de correlación para datos de días no hábiles de Troncal 1



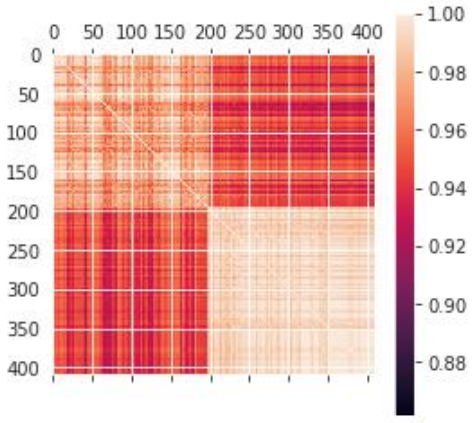
Fuente: Elaboración propia

Ilustración 14. Gráfica de coeficientes de correlación para todos los datos de Troncal 2



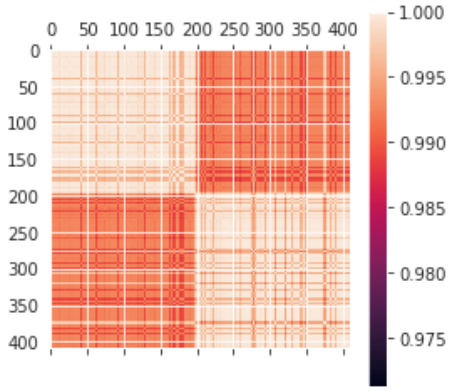
Fuente: Elaboración propia

Ilustración 15. Gráfica de coeficientes de correlación para datos de días hábiles de Troncal 2



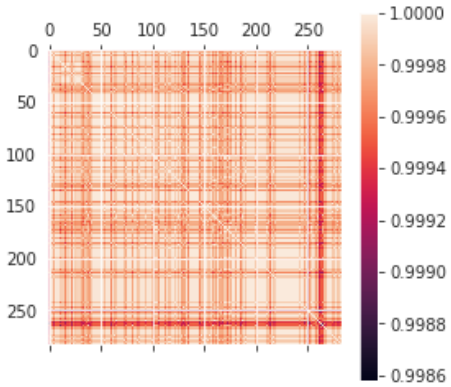
Fuente: Elaboración propia

Ilustración 16. Gráfica de coeficientes de correlación para datos de días no hábiles de Troncal 2



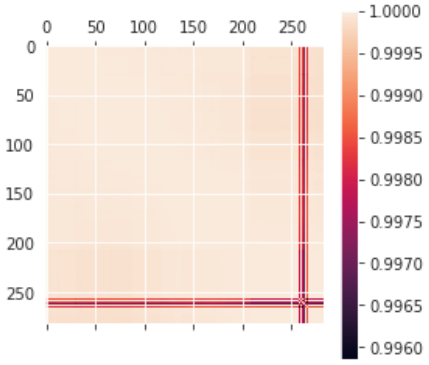
Fuente: Elaboración propia

Ilustración 17. Gráfica de coeficientes de correlación para todos los datos de Troncal 3



Fuente: Elaboración propia

Ilustración 18. Gráfica de coeficientes de correlación para datos de días hábiles de Troncal 3



Fuente: Elaboración propia

En las gráficas de los coeficientes de correlación se observa una fuerte correlación entre los tiempos de llegada a cada una de las geocercas de las troncales (los coeficientes más bajos rondan alrededor de 0.75). Asimismo, se observa que en el caso de las Troncales 1 y 2, se

presenta una estructura que sugiere que, aunque todas las celdas poseen una alta fuerza de asociación, el conjunto de las geocercas que conforman cada sentido del recorrido muestran mayor asociación. De manera general también se evidencia una mayor correlación entre los tiempos de llegada de las geocercas espacialmente más cercanas.

Las matrices de correlación pueden arrojar información valiosa para la identificación de las geocercas que realmente están relacionadas en términos de los tiempos de llegada de los buses. Teniendo esto en cuenta, se considera que en investigaciones futuras se podrían analizar previamente las estructuras de estas matrices, con el objetivo de minimizar el ruido causado por la inclusión de información de estaciones que no estén altamente relacionadas al momento de predecir los tiempos de llegada en una estación particular, sabiendo que existe un radio de influencia espacial.

Una vez analizada la correlación de las variables, se procedió a la estimación de los diversos modelos listados previamente. Estos fueron estimados utilizando la información proveniente de los registros de los GPS incorporados en los buses del sistema entre el 17 y el 30 de marzo del 2014.

Para los modelos de regresión lineal estimados usando las variables explicativas propuestas, se obtuvieron los siguientes coeficientes y respectivo estadístico t:

Tabla 8. Coeficientes obtenidos para los modelos de regresión lineal con variables explicativas para días hábiles

Variables	**Distancia hasta el siguiente paradero aguas abajo**	**Diferencia de headway**	**Diferencia de headway (sin tiempos promedio de parada)**	**Promedio ponderado del tiempo de llegada de buses anteriores**	**Hora pico**	**Intercepto**
Modelo	**Coeficiente V1**	**Coeficiente V2**	**Coeficiente V3**	**Coeficiente V4**	**Coeficiente V5**	
1	0.0362(237.63)	-	3.61E-05(6.07)	0.6535(569.33)	0.9405(16.20)	3.38(77.91)
2	0.0361(237.33)	-	3.17E-05(5.33)	0.6539(569.72)	-	3.62(89.15)
3	0.0361(237.30)	-	-	0.6541(569.99)	-	3.67(92.96)
4	-	-	-	0.8680(1175.34)	-	6.18(156.05)
5	0.0362(237.61)	3.09E-05(5.03)	-	0.6536(569.38)	0.9357(16.13)	3.39(78.51)
6	0.0361(237.32)	2.71E-05(4.41)	-	0.6540(569.77)	-	3.63(89.72)

Fuente: Elaboración propia

Tabla 9. Coeficientes obtenidos para los modelos de regresión lineal con variables explicativas para días no hábiles

Variables	Distancia hasta el siguiente paradero aguas abajo	Diferencia de headway	Diferencia de headway (sin tiempos promedio de parada)	Promedio ponderado del tiempo de llegada de buses anteriores	Hora pico	Intercepto
Modelo	**Coeficiente V1**	**Coeficiente V2**	**Coeficiente V3**	**Coeficiente V4**	**Coeficiente V5**	
1	0.0422(127.16)	-	0.0001(7.68)	0.5855(227.35)	-0.0172(-0.13)	4.40 (48.62)
2	0.0422(127.16)	-	0.0001(7.70)	0.5855(227.35)	-	4.39(51.55)
3	0.0422(127.08)	-	-	0.5863(227.83)	-	4.55(54.90)
4	-	-	-	0.8478(522.82)	-	6.89(81.34)
5	0.0422(127.13)	0.0001(7.35)	-	0.5855(227.38)	-0.0162(-0.26)	4.41(48.86)
6	0.0422(127.14)	0.0001(7.37)	-	0.5855(227.39)	-	4.40(51.83)

Fuente: Elaboración propia

A continuación, se presentan los resultados de los coeficientes de determinación R^2 obtenidos para los diferentes modelos estimados que tuvieron en cuenta las variables explicativas. Las variables consideradas en la estimación de cada uno de los 6 modelos listados en las siguientes tablas están especificadas en la Tabla 6 de la sección 3.4 se distinguen los resultados obtenidos dependiendo de si se consideró la información de los días hábiles o de los días no hábiles:

Tabla 10. Resultados R^2 obtenidos de modelos con variables explicativas para días hábiles

Días Hábiles	Método		
Modelos	Regresión Lineal	ANN	SVM
1	0.69	0.64	0.25
2	0.69	0.65	0.24
3	0.69	0.69	0.65
4	0.67	0.67	0.64
5	0.69	0.57	0.64
6	0.69	0.61	0.28

Fuente: Elaboración propia

Tabla 11. Resultados R^2 obtenidos de modelos con variables explicativas para días no hábiles

No Días Hábiles	Método		
Modelos	Regresión Lineal	ANN	SVM

1	0.67	0.64	0.28
2	0.67	0.65	0.27
3	0.66	0.67	0.62
4	0.63	0.64	0.63
5	0.67	0.66	0.27
6	0.67	0.65	0.26

Fuente: Elaboración propia

Con los resultados obtenidos, se puede evidenciar que el modelo número 3 de los días hábiles ajustado con el modelo ANN arrojó el mejor coeficientes de determinación R^2 de este tipo de modelos. Las variables incluidas en el mismo fueron la distancia hasta el siguiente paradero aguas abajo y promedio ponderado del tiempo de llegada de buses anteriores, las cuales presentaron la mayor correlación con el tiempo de llegada de los buses a las estaciones, como se había mencionado anteriormente.

Al comparar los resultados obtenidos en la estimación de dichos modelos, se puede observar que los mejores ajustes se obtuvieron en aquellos modelos que consideraron menos variables explicativas. Este fenómenos se evidenció principalmente al momento de estimar los modelos con SVM, mostrando que al incluir las variables que no presentaban una fuerte correlación con el tiempo de llegada, genera un ajuste pobre del modelo.

A continuación, se presentan los resultados de los coeficientes de determinación R^2 obtenidos para los modelos que sólo emplearon información histórica de los tiempos de llegada:

Tabla 12. Resultados obtenidos de modelos sólo con información histórica para Troncal 1

Troncal 1			
Método	**Día Hábil**	**Día No Hábil**	**Todos los días**
Lasso	0.933	0.945	0.937
Rigde	0.952	0.931	0.949
kNN	0.950	0.890	0.938

Fuente: Elaboración propia

Tabla 13. Resultados obtenidos de modelos sólo con información histórica para Troncal 2

Troncal 2			
Método	**Día Hábil**	**Día No Hábil**	**Todos los días**

Lasso	0.827	0.974	0.864
Rigde	0.844	0.968	0.876
kNN	0.950	0.974	0.957

Fuente: Elaboración propia

Tabla 14. Resultados obtenidos de modelos sólo con información histórica para Troncal 3

Troncal 3			
Método	**Día Hábil**	**Día No Hábil**	**Todos los días**
Lasso	0.975	-	0.987
Rigde	0.990	-	0.991
kNN	0.918	-	0.993

Fuente: Elaboración propia

Tabla 15. Resultados del parámetro *k* datos más cercanos para los modelos estimados tipo k-NN

k-NN			
Ruta	**Día Hábil**	**Día No Hábil**	**Día completo**
Troncal 1	100	50	100
Troncal 2	50	1	50
Troncal 3	10	-	1

Fuente: Elaboración propia

Es importante aclarar que en este último caso no se estimaron los modelos para los días no hábiles, debido a que, en la información suministrada del mes de junio, no se evidenció que ningún bus de Asemtur realizara la operación de la Troncal 3 por lo que no iba a ser posible realizar la validación de dichas estimaciones.

Se observa que el comportamiento de los 3 modelos es similar, incluso entre las distintas rutas estudiadas, donde se evidencia que no existe una tendencia clara, ya que, dependiendo del tipo de día y ruta estudiada, fue variando el tipo de modelo que resultó ligeramente mejor que los demás.

Además, los resultados sugieren que para el contexto estudiado es posible mezclar los datos de los días hábiles y no hábiles y obtener muy buenos resultados, incluso mejores en algunos casos que los encontrados considerando la información segmentada. Esto puede deberse a que, una vez que se segmenta la información, se cuenta con menor cantidad de información histórica de cada tipo, lo que sugiere que este tipo de modelos podría ser útil en el caso de

que se quiera utilizar toda la información disponible de manera conjunta o se cuente con menos información histórica, caso en el que podría no ser conveniente la segmentación.

Se observó que los modelos donde sólo se consideró el tiempo histórico de llegada a las estaciones presentaron un mejor desempeño que aquellos que se basaron en variables explicativas para realizar la predicción.

Una vez estimados todos los modelos anteriormente presentados, se escogió el modelo número 3 de los días hábiles ajustado con el modelo ANN para ser comparado con los modelos basados únicamente en tiempo mediante las medidas de desempeño de la validación, ya que éste obtuvo el mejor ajuste en los modelos estimados a partir de las variables explicativas.

Para realizar la validación de los modelos que se listaran a continuación se empleó la información proveniente de los registros de los GPS incorporados en los buses del sistema entre el 1 y el 6 de junio del 2014. Los indicadores utilizados para poder comparar los resultados de las validaciones de los distintos modelos son el EAM, EPAM y RECM, los cuales fueron detallados en la sección 3.5. Estos fueron calculados considerando la información estimada utilizando los métodos propuestos y la información real disponible. A continuación, se resumen los resultados:

Tabla 16. Medidas de desempeño del mejor modelo con variables explicativas

Método: ANN - Modelo 3 Días hábiles	
EAM	14.841
EPAM	43.38%
RECM	23.082

Fuente: Elaboración propia

Tabla 17. EAM para modelos sólo con información histórica para Troncal 1

Troncal 1			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	0.850	0.872	0.834
Rigde	1.047	0.704	0.706
k-NN	0.736	1.054	0.775
Inferencia bayesiana	0.036	0.008	0.036

Fuente: Elaboración propia

Tabla 18. EAM para modelos sólo con información histórica para Troncal 2

Troncal 2			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	0.883	0.585	0.815
Rigde	0.860	0.640	0.684
k-NN	0.720	0.622	0.684
Inferencia bayesiana	0.020	0.025	0.021

Fuente: Elaboración propia

Tabla 19. EAM para modelos sólo con información histórica para Troncal 3

Troncal 3			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	0.662	-	0.482
Rigde	0.461	-	0.444
k-NN	1.064	-	0.367
Inferencia bayesiana	0.007	-	0.014

Fuente: Elaboración propia

Tabla 20. EPAM para modelos sólo con información histórica para Troncal 1

Troncal 1			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	8.09%	7.12%	7.60%
Rigde	10.88%	5.60%	6.47%
k-NN	6.29%	7.84%	6.26%
Inferencia bayesiana	0.39%	0.06%	0.40%

Fuente: Elaboración propia

Tabla 21. EPAM para modelos sólo con información histórica para Troncal 2

Troncal 2			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	7.60%	5.03%	7.02%
Rigde	6.94%	4.99%	6.06%
k-NN	6.53%	4.92%	6.06%
Inferencia bayesiana	0.23%	0.26%	0.23%

Fuente: Elaboración propia

Tabla 22. EPAM para modelos sólo con información histórica para Troncal 3

Troncal 3

Método	Día Hábil	Día No Hábil	Día completo
Lasso	3.90%	-	3.80%
Rigde	3.90%	-	3.80%
k-NN	9.85%	-	3.25%
Inferencia bayesiana	0.060%	-	0.110%

Fuente: Elaboración propia

Tabla 23. RECM para modelos sólo con información histórica para Troncal 1

Troncal 1			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	1.333	1.251	1.267
Rigde	2.748	1.120	1.144
k-NN	1.150	1.583	1.261
Inferencia bayesiana	0.189	0.092	0.189

Fuente: Elaboración propia

Tabla 24. RECM para modelos sólo con información histórica para Troncal 2

Troncal 2			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	3.647	0.581	2.959
Rigde	3.291	0.706	0.932
k-NN	1.059	0.573	0.932
Inferencia bayesiana	0.092	0.086	0.083

Fuente: Elaboración propia

Tabla 25. RECM para modelos sólo con información histórica para Troncal 3

Troncal 3			
Método	**Día Hábil**	**Día No Hábil**	**Día completo**
Lasso	0.546	-	0.273
Rigde	0.220	-	0.204
k-NN	1.797	-	0.158
Inferencia bayesiana	0.010	-	0.032

Fuente: Elaboración propia

Como se observa en la Tabla 16, el modelo número 3 de los días hábiles ajustado con el modelo ANN obtuvo un 43.38% en el EPAM, mientras que el modelo con el ajuste más bajo que sólo tuvo en cuenta la variable tiempo, el cual fue la regresión Lasso para modelo de los

días hábiles de la Troncal 2, en la Tabla 21 se observa que presentó un 7.60% en el EPAM, lo que nos lleva a concluir que, para el presente caso de estudio y condiciones operacionales, este tipo de modelos son mucho más adecuados. Asimismo, se observa que el comportamiento del modelo de inferencia bayesiana muestra excelentes resultados, debido a que todos presentaron EPAM por debajo al 1%.

Este tipo de resultados deja en evidencia que para la implementación de modelos de predicción del tiempo de llegada mediante variables explicativas, es fundamental que éstos cuenten con variables muy bien especificadas, adecuadas y con una buena medición, así como con parámetros en los modelos de predicción bien ajustados a la información disponible, como por ejemplo escoger el número de capas o neuronas adecuado, funciones de activación y algoritmos de solución que brinden el mejor desempeño, dado que los factores anteriormente mencionados pueden introducir ruido a las estimaciones e inducir a predicciones erróneas.

Asimismo, se realizó una estimación del modelo bayesiano propuesto en un contexto online. En este caso se obtienen registros GPS de los buses en tiempo real (a medida que el bus se va moviendo); cada dato nuevo registrado es incorporado a la base de datos histórica que se usa como prior para la predicción del tiempo de llegada al punto siguiente. Dado que el bus eventualmente llegará al punto para el cual se acaba de realizar la predicción, es posible realizar una validación entre lo que se predijo y la información real medida por el GPS. Este tipo de metodología online es provechosa porque se alimenta tanto de la distribución de los datos históricos como de lo que el mismo bus viene reflejando en su recorrido actual. Algo importante para la implementación de este tipo de predicción online es que los tiempos computacionales sean bajos, de lo contrario no se podrán tener resultados en tiempo real. El modelo online presentado en esta tesis es rápido, eficiente y arroja mejores resultados por lo que resulta ser la mejor opción para implementar en el caso de que se tenga la estructura de recolección y transmisión de datos necesaria en el sistema.

Se hace claridad que los modelos anteriormente presentados fueron estimados en un contexto offline, es decir, que la información del mes de junio no fue incorporada para actualizar el modelo estimado, fue usada sólo para validar los modelos.

Para la estimación en el contexto online, la información del mes de marzo fue empleada como la *prior* inicial del modelo, mientras que la información del mes de junio fue utilizada para validar y alimentar la *prior* con la información en tiempo real recibida. Esto lo convierte en un modelo dinámico, que no sólo emplea la información histórica, sino que tiene en cuenta las condiciones actuales de la red por la nueva información incorporada.

Los indicadores utilizados para poder comparar los resultados de las validaciones de los distintos modelos son el EAM, EPAM y RECM, los cuales fueron detallados en la sección 3.5. Estos fueron calculados considerando la información estimada utilizando los métodos propuestos y la información real disponible. A continuación, se resumen los resultados:

Tabla 26. EAM para modelos de inferencia bayesiana en contexto online

Ruta	Día Hábil	Día No Hábil	Día completo
Troncal 1	0.034	0.029	0.032
Troncal 2	0.021	0.023	0.021
Troncal 3	0.015	-	0.013

Fuente: Elaboración propia

Tabla 27. EPAM para modelos de inferencia bayesiana en contexto online

Ruta	Día Hábil	Día No Hábil	Día completo
Troncal 1	0.37%	0.30%	0.35%
Troncal 2	0.24%	0.23%	0.23%
Troncal 3	0.12%	-	0.10%

Fuente: Elaboración propia

Tabla 28. RECM para modelos de inferencia bayesiana en contexto online

Ruta	Día Hábil	Día No Hábil	Día completo
Troncal 1	0.135	0.114	0.134
Troncal 2	0.094	0.080	0.081
Troncal 3	0.026	-	0.021

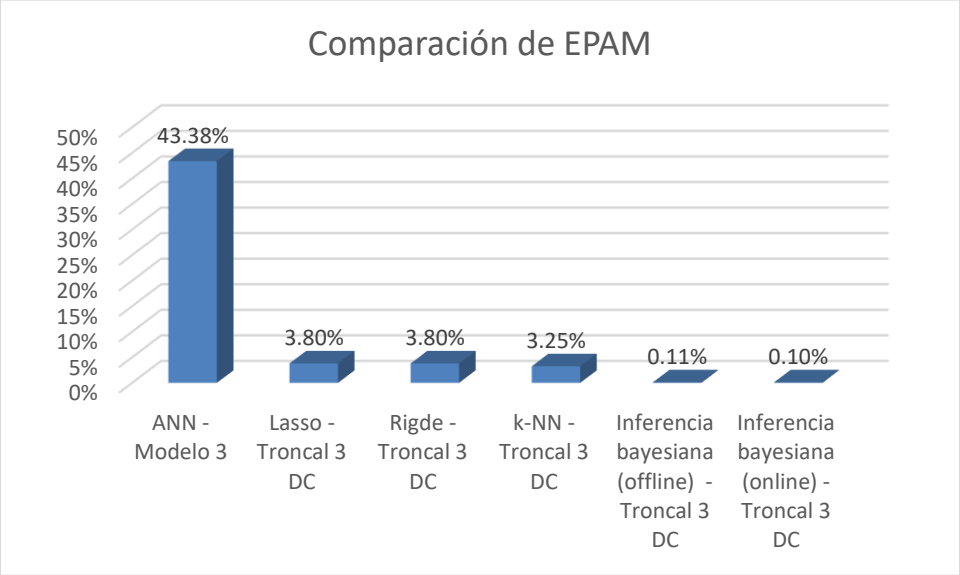
Fuente: Elaboración propia

Como se observa en la Tabla 27, para la mayoría de los casos evaluados, los resultados obtenidos para el EPAM del modelo de inferencia bayesiana en el contexto online presentan una mejora con respecto al mismo modelo estimado en un contexto offline, disminuyendo este indicador hasta en un 0.05%. Aunque, claramente el modelo de inferencia bayesiana en el contexto offline presenta excelentes resultados, al migrar éste a un contexto online, se enriquece la predicción realizada, ya que refleja el estado actual del recorrido del bus.

Por último, acorde con los resultados obtenidos en la validación de los distintos modelos, se puede concluir que el método de la inferencia bayesiana fue el que mejor desempeño presentó, con un EPAM promedio del 0.24%. Demostrando la utilidad sobre emplear una función de probabilidad que permita describir la dinámica de los datos históricos, que en este caso es el tiempo de llegada de los bus.

A continuación, se presenta una gráfica comparativa de los modelos y las medidas de desempeño calculadas para plasmar el desempeño de cada uno:

Ilustración 19. Comparación de resultados de EPAM de todos los modelos



Fuente: Elaboración propia

Así las cosas, y revisando los resultados obtenidos de los diferentes modelos estimados, se puede concluir que, para el caso de estudio evaluado, teniendo las mismas condiciones operacionales, sólo basándonos en la variable del tiempo de llegada del bus, se obtienen resultados sustancialmente mejores, con modelos que presentan soluciones altamente eficientes y computacionalmente viables, sin la necesidad de calcular tantas variables explicativas.

Es importante destacar además que, una de las ventajas que presenta el modelo bayesiano sobre los otros modelos presentados es su dinamismo, por medio del cual permite la actualización de la información en tiempo real, a medida que se reciben nuevos registros del GPS del bus, lo que hace que sea más susceptible a captar y adaptarse mejor algún evento

atípico presentado en el sistema, ya sea por algún problema operacional o un incidente de tránsito que afecte a los buses, no como los otros modelos de predicción que deben ser reentrenados con información histórica cada tanto.

5. CONCLUSIONES

Hoy en día proporcionar información precisa sobre los horarios de llegada de los buses en las paradas es uno de los parámetros clave para el ofrecimiento de un servicio de transporte público alta calidad. A pesar de que se han desarrollado numerosos programas y sistemas para predecir este tipo de información en tiempo real, aún sigue siendo una labor compleja, y existen desviaciones entre las predicciones obtenidas y el tiempo de viaje real. Estas desviaciones suelen deberse a varios factores estocásticos, incluido el comportamiento del conductor, el ancho de la vía del carril, las intersecciones, la congestión, la demanda de viajes, la señalización, el clima, entre otros. En la revisión del estado del arte se pudo evidenciar que estos factores se utilizan como variables independientes explicativas en muchos de los modelos formulados. La precisión de estos métodos depende de la correcta estimación y asociación de todas las variables que sean reconocidas e incorporadas en el modelo, lo cual es un procedimiento difícil. A lo largo de los años, los investigadores han buscado reducir en la mayor proporción posible las desviaciones proponiendo gran variedad de modelos ya que la información que se estime debe ser relevante y confiable; de lo contrario su utilización podría tener un efecto negativo, tanto en la operación como en la percepción de los usuarios.

En la presente investigación se formularon una serie de modelos basados en datos históricos y en tiempo real en algunos casos para la predicción de los tiempos de llegada de los buses a las estaciones. En este caso se tomaron como caso de estudio las estaciones del sistema troncal de Megabus, sistema tipo BRT que opera en la ciudad de Pereira en Colombia. Las metodologías empleadas incluyeron enfoques desde los más simples hasta los más complejos: velocidad promedio, regresión lineal, redes neuronales artificiales (ANN), máquina de vectores de soporte (SVM), regresión Ridge, regresión Lasso y además un método bayesiano propuesto para el desarrollo de esta tesis que considera la actualización de la información histórica (caracterizada por una distribución de probabilidad) a medida que se obtiene nueva información por parte de los GPS de los buses. Uno de los aspectos que se quería evaluar que variables era más eficiente utilizar para realizar la predicción. Dependiendo de la metodología empleada los modelos fueron estimados utilizando información sobre la distancia hasta la parada, velocidad, diferencia de headway, promedio

ponderado del tiempo de llegada de buses anteriores, el efecto de la hora pico y/o el tiempo de llegada. Los resultados indican que la utilización de mayor número de variables no necesariamente implica un mejor ajuste del modelo. En este caso, al realizar el análisis y la comparación los modelos estimados, se concluyó que el modelo del propuesto que dependen únicamente de la variable de tiempo de llegada con el método de la inferencia bayesiana en el contexto online, fue el que mejor desempeño presentó, con un EPAM promedio del 0.24%, seguido por el método de la inferencia bayesiana en el contexto offline, y luego los modelos seguido de los modelos de regresión Ridge, Lasso y kNN.

Como el propósito de la presente investigación no es entender el efecto de las variables independientes sobre la hora de llegada de los buses a las estaciones sino obtener una predicción acertada de dicha variable, tanto los métodos de regresión por Ridge, Lasso y kNN, como el método por inferencia bayesiana propuesto que dependen únicamente de la variable de tiempo de llegada se pueden considerar soluciones altamente eficientes y computacionalmente viables para la predicción de la hora de llegada de los buses a las paradas en tiempo real. En el caso de la inferencia bayesiana, el modelo además es dinámico y permite la actualización de la información a medida que se reciben nuevos registros del GPS del bus.

También es importante destacar si se busca realizar la implementación de un modelo de predicción mediante variables explicativas, se debe tener en cuenta que hay diversos factores que pueden inducir a errores en dichas estimaciones, como por ejemplo: considerar variables que realmente no influyan en el modelo explicativo del tiempo de llegada, realizar una mala medición de la variable, tener una base de datos con pocas observaciones, escoger parámetros en los modelos de predicción que no se ajusten a la información disponible (al momento de escoger el número de capas, funciones de activación, algoritmos para resolver, etc). Por tal motivo, se sugiere realizar las comparaciones con modelos que cuenten sólo con información histórica y realizar la comparación para cada contexto evaluado.

Asimismo, es importante tener en cuenta que si se implementa algún modelo estimado en un contexto offline (sólo con información histórica), y las condiciones operacionales del sistema presentan un cambio significativo, se debe recolectar nuevos datos y volver a realizar la estimación del modelo con el nuevo escenario. Así las cosas, se debe tener en cuenta que este

tipo de modelos se debe reentrenar cada cierto tiempo, dependiendo diferentes características del sistema de transporte en que se aplique, tales como flujos de vehículos, mes del año, si cuenta con un carril exclusivo para los buses, entro otras.

Se analizaron además las matrices de correlación de las predicciones entre las estaciones del sistema y se evidenciaron patrones marcados que sugieren que existe una especie de zona de influencia. Esto indica que el tiempo de llegada de un bus a una determinada estación está fuertemente correlacionado únicamente con los tiempos de llegada de las estaciones más cercanas. En investigaciones futuras se podría hacer un análisis de las estructuras de las matrices de correlación para involucrar en las predicciones únicamente información proveniente de estaciones que estén fuertemente correlacionadas con la estación para la cual se esté realizando la predicción, evitando así que ingrese ruido innecesario a la estimación de estaciones que no tienen influencia y se puedan obtener resultados aún más precisos. Estas estructuras de correlación dependerán mucho de la configuración de cada sistema y de sus características de operación.

Es importante mencionar que, si bien en el contexto estudiado los resultados del modelo bayesiano propuesto son superiores a los encontrados en la literatura, puede que en otras situaciones en las que los tiempos de viaje de los buses sean altamente variables y fuertemente influenciados por factores estocásticos externos sea necesario considerar la formulación de otro tipo de modelos que tengan en cuenta mayor cantidad de variables explicativas.

6. REFERENCIAS

Altinkaya, M., & Zontul, M. (2013). Urban Bus Arrival Time Prediction: A Review of Computational Models. *International Journal of Recent Technology and Engineering*, *2*(4), 2277–3878.

Arellana, J., De Dios Ortúzar, J., Rizzi, L. I., & Zuñiga, F. (2014). Obtaining public transport level-of-service measures using in-vehicle GPS data and freely available GIS web-based tools. *Mobile Technologies for Activity-Travel Data Collection and Analysis*, *January*, 258–275. https://doi.org/10.4018/978-1-4666-6170-7.ch016

Bae, S. (1997). Dynamic estimation of travel time on arterial roads by using automatic vehicle location (AVL) bus as a vehicle probe. *Transportation Research Part A: Policy and Practice*, *31*(1), 60. https://doi.org/10.1016/s0965-8564(97)88279-1

Basso, L. J., Feres, F., & Silva, H. E. (2019). The efficiency of bus rapid transit (BRT) systems: A dynamic congestion approach. *Transportation Research Part B: Methodological*, *127*(December 2018), 47–71. https://doi.org/10.1016/j.trb.2019.06.012

Bin, Y., Zhongzhen, Y., & Baozhen, Y. (2006). Bus arrival time prediction using support vector machines. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *10*(4), 151–158. https://doi.org/10.1080/15472450600981009

Cathey, F. W., & Dailey, D. J. (2003). A prescription for transit arrival/departure prediction using automatic vehicle location data. *Transportation Research Part C: Emerging Technologies*, *11*(3–4), 241–264. https://doi.org/10.1016/S0968-090X(03)00023-8

Čelan, M., & Lep, M. (2017). Bus arrival time prediction based on network model. *Procedia Computer Science*, *113*, 138–145. https://doi.org/10.1016/j.procs.2017.08.331

Čelan, M., & Lep, M. (2020). Bus-arrival time prediction using bus network data model

and time periods. *Future Generation Computer Systems*, *110*, 364–371. https://doi.org/10.1016/j.future.2018.04.077

Chen, M., Liu, X., Xia, J., & Chien, S. I. (2004). A dynamic bus-arrival time prediction model based on APC data. *Computer-Aided Civil and Infrastructure Engineering*, *19*(5), 364–376. https://doi.org/10.1111/j.1467-8667.2004.00363.x

Chen, X. M., Gong, H. B., & Wang, J. N. (2012). BRT vehicle travel time prediction based on SVM and Kalman filter. *Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology*, *12*(4), 29–34. https://doi.org/10.1016/s1570-6672(11)60211-0

Chien, S. I. J., Ding, Y., & Wei, C. (2002). Dynamic bus arrival time prediction with artificial neural networks. *Journal of Transportation Engineering*, *128*(5), 429–438. https://doi.org/10.1061/(ASCE)0733-947X(2002)128:5(429)

Chien, S. I. J., & Kuchipudi, C. M. (2003). Dynamic travel time prediction with real-time and historic data. *Journal of Transportation Engineering*, *129*(6), 608–616. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(608)

Comi, A., Zhuk, M., Kovalyshyn, V., & Hilevych, V. (2020). Investigating bus travel time and predictive models: A time series-based approach. In *Transportation Research Procedia* (Vol. 45, pp. 692–699). https://doi.org/10.1016/j.trpro.2020.02.109

Cortés, C. E., Gibson, J., Gschwender, A., Munizaga, M., & Zúñiga, M. (2011). Commercial bus speed diagnosis based on GPS-monitored data. *Transportation Research Part C: Emerging Technologies*, *19*(4), 695–707. https://doi.org/10.1016/j.trc.2010.12.008

D'Angelo, M. P., Al-Deek, H. M., & Wang, M. C. (1999). Travel-time prediction for freeway corridors. *Transportation Research Record*, *1676*, 184–191. https://doi.org/10.3141/1676-23

de Grange, L., Fariña, P., & de Dios Ortúzar, J. (2015). Dealing with collinearity in travel time valuation. *Transportmetrica A: Transport Science*, *11*(4), 317–332.

https://doi.org/10.1080/23249935.2014.988195

Dell'Olio, L., Ibeas, A., & Cecin, P. (2011). The quality of service desired by public transport users. *Transport Policy*, *18*(1), 217–227. https://doi.org/10.1016/j.tranpol.2010.08.005

Deng, L., He, Z., & Zhong, R. (2013). The bus travel time prediction based on Bayesian networks. *Proceedings - 2013 International Conference on Information Technology and Applications, ITA 2013*, 282–285. https://doi.org/10.1109/ITA.2013.73

Dhivya Bharathi, B., Anil Kumar, B., Achar, A., & Vanajakshi, L. (2020). Bus travel time prediction: a log-normal auto-regressive (AR) modelling approach. *Transportmetrica A: Transport Science*, *16*(3), 807–839. https://doi.org/10.1080/23249935.2020.1720864

Dziekan, K. (2008). Ease-of-use in public transportation: a user perspective on information and orientation aspects. In *Department of Transport and Economics*.

Elragal, A., & Raslan, H. (2014). Analysis of trajectory data in support of traffic management: A data mining approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8557 LNAI*, 174–188. https://doi.org/10.1007/978-3-319-08976-8_13

Gurmu, Z. K., & Fan, W. D. (2014). Artificial neural network travel time prediction model for buses using only GPS data. *Journal of Public Transportation*, *17*(2), 45–65. https://doi.org/10.5038/2375-0901.17.2.3

Hastie, T. et. all. (2009). Springer Series in Statistics The Elements of Statistical Learning. *The Mathematical Intelligencer*, *27*(2), 83–85. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Haworth, J., Shawe-Taylor, J., Cheng, T., & Wang, J. (2014). Local online kernel ridge regression for forecasting of urban travel times. *Transportation Research Part C: Emerging Technologies*, *46*, 151–178. https://doi.org/10.1016/j.trc.2014.05.015

He, P., Jiang, G., Lam, S. K., & Sun, Y. (2020). Learning heterogeneous traffic patterns for travel time prediction of bus journeys. *Information Sciences*, *512*, 1394–1406. https://doi.org/10.1016/j.ins.2019.10.073

Hou, Y., & Edara, P. (2018). Network Scale Travel Time Prediction using Deep Learning. *Transportation Research Record*, *2672*(45), 115–123. https://doi.org/10.1177/0361198118776139

Jeong, R., & Rilett, L. (2004). The prediction of Bus Arrival Time using AVL data. In *83rd Annual General Meeting, Transportation Research Board, National Research Council, Washington DC, USA.*

Kamarianakis, Y., Shen, W., & Wynter, L. (2012). Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. *Applied Stochastic Models in Business and Industry*, *28*(4), 297–315. https://doi.org/10.1002/asmb.1937

Kumar, B. A., Vanajakshi, L., & Subramanian, S. C. (2017). Bus travel time prediction using a time-space discretization approach. *Transportation Research Part C: Emerging Technologies*, *79*, 308–332. https://doi.org/10.1016/j.trc.2017.04.002

Lin, W. H., & Zeng, J. (1999). Experimental study of real-time bus arrival time prediction with GPS data. *Transportation Research Record*, *1666*, 101–109. https://doi.org/10.3141/1666-12

Ma, J., Chan, J., Ristanoski, G., Rajasegarar, S., & Leckie, C. (2019). Bus travel time prediction with real-time traffic information. *Transportation Research Part C: Emerging Technologies*, *105*(June 2018), 536–549. https://doi.org/10.1016/j.trc.2019.06.008

Madigan, D. (2008). Least Angle Regression. *The Science of Bradley Efron*, 385–479. https://doi.org/10.1007/978-0-387-75692-9_20

Mori, U., Mendiburu, A., Álvarez, M., & Lozano, J. A. (2015). A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science*, *11*(2), 119–157.

https://doi.org/10.1080/23249935.2014.932469

Patnaik, J., Chien, S., & Bladikas, A. (2004). Estimation of Bus Arrival Times Using APC Data. *Journal of Public Transportation*, *7*(1), 1–20. https://doi.org/10.5038/2375-0901.7.1.1

Philip, A. M., Ramadurai, G., & Vanajakshi, L. (2018). Urban Arterial Travel Time Prediction Using Support Vector Regression. *Transportation in Developing Economies*, *4*(1), 1–8. https://doi.org/10.1007/s40890-018-0060-6

Ravada, S., Ali, M., Bao, J., & Sarwat, M. (2013). International Conference on Advances in Geographic Information Systems Cup 2013: Geo-fencing. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 574–577.

Sun, D., Luo, H., Fu, L., Liu, W., Liao, X., & Zhao, M. (2007). Predicting bus arrival time on the basis of global positioning system data. *Transportation Research Record, 2034*, 62–72. https://doi.org/10.3141/2034-08

Tang, K., Chen, S., Khattak, A. J., & Pan, Y. (2019). Deep Architecture for Citywide Travel Time Estimation Incorporating Contextual Information. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *0*(0), 1–17. https://doi.org/10.1080/15472450.2019.1617141

Van lint, J. W. . (2004). *Reliable Travel Time Prediction for Freeways*.

Vanajakshi, L., & Rilett, L. R. (2007). Support vector machine technique for the short term prediction of travel time. *IEEE Intelligent Vehicles Symposium, Proceedings*, 600–605. https://doi.org/10.1109/ivs.2007.4290181

Wang, L., Zuo, Z., & Fu, J. (2014). Bus Arrival Time Prediction Using RBF Neural Networks Adjusted by Online Data. *Procedia - Social and Behavioral Sciences*, *138*(0), 67–75. https://doi.org/10.1016/j.sbspro.2014.07.182

Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-time prediction with support vector

regression. *IEEE Transactions on Intelligent Transportation Systems*, *5*(4), 276–281. https://doi.org/10.1109/TITS.2004.837813

Xin, J., & Chen, S. (2016). Bus Dwell Time Prediction Based on KNN. *Procedia Engineering*, *137*, 283–288. https://doi.org/10.1016/j.proeng.2016.01.260

Xu, M., Guo, K., Fang, J., & Chen, Z. (2019). Utilizing Artificial Neural Network in GPS-Equipped Probe Vehicles Data- Based Travel Time Estimation. *IEEE Access*, *7*, 89412–89426. https://doi.org/10.1109/ACCESS.2019.2926851

Yu, B., Lam, W. H. K., & Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, *19*(6), 1157–1170. https://doi.org/10.1016/j.trc.2011.01.003

Yu, H., Chen, D., Wu, Z., Ma, X., & Wang, Y. (2016). Headway-based bus bunching prediction using transit smart card data. *Transportation Research Part C: Emerging Technologies*, *72*, 45–59. https://doi.org/10.1016/j.trc.2016.09.007

Yu, Z., Wood, J. S., & Gayah, V. V. (2017). Using survival models to estimate bus travel times and associated uncertainties. *Transportation Research Part C: Emerging Technologies*, *74*, 366–382. https://doi.org/10.1016/j.trc.2016.11.013

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zheng, C. J., Zhang, Y. H., & Feng, X. J. (2012). Improved iterative prediction for multiple stop arrival time using a support vector machine. *Transport*, *27*(2), 158–164. https://doi.org/10.3846/16484142.2012.692710

Zhou, P., Zheng, Y., & Li, M. (2014). How long to wait? Predicting bus arrival time with mobile phone based participatory sensing. *IEEE Transactions on Mobile Computing*, *13*(6), 1228–1241. https://doi.org/10.1109/TMC.2013.136